DOCTORAL THESIS

Automating Systematic Reviews

Rathbone, John

Award date:
2017

Automating Systematic Reviews

John Rathbone

A thesis submitted in total fulfilment of the requirement of the degree of
Doctor of Philosophy (PhD)

May 2017

Centre for Research in Evidence-Based Practice
Faculty of Health Science and Medicine

Professors Paul Glasziou, Tammy Hoffmann & Associate Professor Elaine Beller

This research was supported by an Australian Government Research Training Program Scholarship.

Abstract

Background

Systematic reviews are used as the 'gold standard' to evaluate healthcare, education, and social policies. They are integral to the clinical decision making of healthcare professionals, and funding decisions made by governmental agencies. The rapid growth in primary research has not been matched by a growth in the efficiency of producing systematic reviews and consequently evidence-based decision making is struggling to remain feasible.

Aims

This body of research aimed to develop and evaluate strategies towards the automation of systematic reviews, so that secondary health research can be produced more efficiently and cost effectively. To that end, four research studies were developed: 1. Comparing the performance of biomedical databases to determine the sensitivity and precision for identifying systematic reviews; 2. Developing and evaluating algorithms to detect duplicate records arising from searching biomedical databases; 3. Evaluating the potential benefits from using a semi-automated machine learning predictive algorithm for citation screening; 4. Developing and evaluating strategies to expedite citation screening using title-only keyword searching.

Methods

Different methods were used to answer the research questions. For the first research study (identifying reviews), 7 biomedical databases were searched for systematic reviews of any intervention for hypertension and the performance of each database was assessed and compared for both comprehensiveness and accuracy. For the second research study (deduplication), an iterative approach was needed to develop and evaluate the performance of each algorithm to detect duplicates; the results acquired from each algorithm were used to inform the next iteration until an ideal algorithm was produced that achieved higher duplicate detection than current methods, but without compromising accuracy. For the third research study (predictive screening), 4 datasets from the literature searches of

published systematic reviews were used to evaluate an online machine learning predictive algorithm by replicating the screening decisions of the original reviews; sensitivity analyses were performed to determine if the reduction in screening effort could be further improved by including non-relevant citations that were closely matched to the review inclusion criteria. For the fourth study (expediting screening), 10 datasets from the literature searches of published systematic reviews were used to evaluate title-only screening. Datasets were screened using title-only keywords searching based upon the inclusion criteria of each systematic review. The results were compared against the published reviews for reduction in screening effort and recall of included studies.

Results

In the first study, the biomedical database, EMBASE, retrieved the largest number of relevant citations (69% sensitivity), but also was the least specific (7% specificity), retrieving many irrelevant citations. The Cochrane Library had 60% sensitivity and was the most precise (30%) of all the databases. None of the databases identified all the relevant records, but a combination of EMBASE, the Cochrane Library and Epistemonikos identified 83% of all the relevant systematic reviews.

In the second study, the iteratively developed deduplication algorithm increased duplicate detection by an average of 42% compared with duplicate detection using EndNote™ bibliographic reference management software. Additionally, all unique citations were correctly classified, whereas EndNote™ classified some unique citations wrongly as duplicate records.

In study 3, the evaluation found that the predictive screening tool (Abstrackr) reduced the screening effort in a range from 9% to 57% depending on the complexity of the systematic review. The reliability to retrieve included studies was good, with most relevant citations found, but in 2 datasets one included study was not retrieved by Abstrackr. Sensitivity analyses found that workload savings could be further increased by including closely matched non-relevant citations, and very large datasets (≥15,000 citations) could achieve as much as 80% reduction in screening.

In study 4, the interest was to reduce screening effort using title-only screening. This ranged from 11% to 78% with a median reduction in screening effort of 53%. In

9 systematic reviews the recall of included studies was 100%. In one review, 4 of 5 reviewers did not identify the same included study (median recall: 67%, total included studies *n*=3).

Discussion and implications

Automation tools are increasingly being developed and interest in the subject continues to grow with new automation methods and literature overviews being published. Some of the automation tools have not been fully tested and this is likely to be a barrier to implementation by systematic reviewers. Other tools show promise but have not been developed into consumer level products. As a response to these challenges, working parties have been established to overcome these barriers and establish a set of principles and goals. The findings from this body of research have shown that more efficient working practices are possible through improved duplicate detection and can be made available to the systematic review community without a prolonged research and development period. The clear potential for machine learning algorithms to automate decisions and reduce screening was demonstrated, but has not been realised into a consumer ready product, and therefore is worthy of further research and development. Biomedical databases offer different products which vary in scale and content and researchers should be prepared to search several databases rather than relying on a single database. The title-only screening developed during this research was shown to be effective and demonstrated similar reliability to both predictive screening tools and human screening, and could be used with other automation tools to assist with screening. Progress with automation tools will be accelerated once technical barriers are overcome, and by pursuing proof of concept technologies into consumer ready products and thoroughly evaluating automation tools for reliability.

Keywords

- Abstrackr
- Algorithm
- Automation
- Biomedical database
- Citation screening
- Deduplication
- Expediting Evidence Synthesis
- Machine learning
- PICO based title-only screening
- Rapid Review
- Scoping Search
- Semi-automation
- Systematic Review Assistant-Deduplication Module
- SRA
- Systematic Review

Declaration by Author

This thesis is submitted to Bond University in fulfilment of the requirements of the degree of *Doctor of Philosophy*. I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three original papers published in peer reviewed journals and one *in press* publication. The core theme of the thesis is developing and investigating methods to increase the efficiency of producing evidence-based medicine research, specifically systematic reviews. The publications and thesis represents my own original work under the supervision of Paul Glasziou, Tammy Hoffmann and Elaine Beller. The inclusion of co-authors is demonstrative of input between multidisciplinary researchers and acknowledges collaboration with team-based research methods.

John Rathbone 31/08/2017

Research outputs

Peer-reviewed publications:

1) Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. **Systematic Reviews (2015)** 4**:6.**

2) Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. **Systematic Reviews (2015) 4:80.**

3) Rathbone J, Carter M, Hoffmann T, Glasziou P. A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. **Systematic Reviews (2016) 5:27.**

4) Rathbone J, Albarqouni L, Bakhit M, Beller E, Byambasuren O, Hoffmann T, Scott AM, Glasziou P. Expediting citation screening using PICo based title-only screening for identifying rapid reviews **Systematic Reviews (2017)** (*In Press*).

Published and Presented Conference Abstracts

5) Rathbone J, Carter M, Hoffmann T, Glasziou P. Solving research waste with better duplicate detection. **October 2015. European Journal of Public Health Conference,** 25 (suppl. 3).

Declaration of authors contribution

Publications co-authored	Statement of contribution
Rathbone, J., M. Carter, T. Hoffmann, P. Glasziou (2015). *Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module.* Syst Rev 4(1)	JR (70%), MC (20%), TH (5%), PG (5%)
Rathbone, J., T. Hoffmann, P. Glasziou (2015). *Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers.* Syst Rev 4(1)	JR (90%), TH (5%), PG (5%)
Rathbone, J., M. Carter, T. Hoffmann, P. Glasziou (2016). A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. Syst Rev 5(27)	JR (80%), MC (10%), TH (5%), PG (5%)
John Rathbone, Loai Albarqouni, Mina Bakhit, Elaine Beller, Oyungerel Byambasuren, Tammy Hoffmann, Anna Mae Scott, Paul Glasziou *(in Press).* Expediting citation screening using PICo based title-only screening for identifying studies in Rapid Reviews	JR (74%), LA (5%), MB (5%), EB (2%), OB (5%), TH (2%), AMS (5%), PG (2%)

Copyright declaration

Acknowledgements

I would like to express my gratitude to my supervisors Paul Glasziou, Tammy Hoffmann, and Elaine Beller for their continuous support and guidance during my PhD and for providing such an enjoyable academic study environment.

I would also like to thank my colleagues at the Centre for Evidenced-Based Practice (CREBP), Bond University for their support and collegiality, and for making the research environment so informal and engaging. Special thanks to Matt Carter whom I collaborated with during my research. In addition, I wish to acknowledge with thanks the Australia Fellowship grant from the National Health and Medical Research Council, and the academic support from the school of Health Sciences and Medicine, Bond University.

My thanks to the external examiners - Hans Lund, Iain Marshall and Karen Robinson, for giving-up so freely their precious time, and for their thoughtful consideration of the thesis.

And to my wife, Evelyne, for the ongoing support and planting the idea that led to uprooting to the antipodes and embarking upon a PhD programme.

Table of Contents

List of Tables

Chapter 3 Biomedical database coverage

Chapter 4 Duplicate detection within bibliographic records

List of Figures

Chapter 1 Introduction

Chapter 3 Biomedical database coverage

Chapter 5 Semi-automated citation screening

Chapter 6 Screening citations using PICo based title-only screening

List of Abbreviations

aHUS – Atypical Haemolytic Uraemic Syndrome

API - Application Programming Interface

CDSR - Cochrane Database of Systematic Reviews

CENTRAL - Cochrane Central Register of Controlled Trials

CINAHL - Cumulative Index to Nursing and Allied Health Literature

CONSORT - Consolidated Standards of Reporting Trials

CREBP – Centre for Research in Evidence based Practice

DARE - Database of Abstracts of Reviews of Effects

DICOM - Digital Imaging and Communications in Medicine

ECHO - Echocardiography

EMBASE - Excerpta Medica Database

EBM – Evidence Based Medicine

FN – False Negative

FP – False Positive

HTA - Health Technology Appraisal

IBM - International Business Machines Corporation

ICASR - International Collaboration for the Automation of Systematic Reviews

LILACS - Latin American and Caribbean Health Sciences Literature

MEDLINE - Medical Literature Analysis and Retrieval System Online

MeSH - Medical Subject Headings

PICO - Participants, Interventions, Comparators, Outcomes

PICo - Participants, Interventions, Comparators, outcomes

PICOS - Participants, Interventions, Comparators, Outcomes, Study design

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analysis

RSE – Reduction in Screening Effort

SE – Screening Effort

SRA-DM – Systematic Review Assistant – Deduplication algorithm

TN – True Negative

TRIP - Turning Research Into Practice

"*In 18th century England, James Hargreaves [an illiterate weaver] invented the Spinning Jenny, and Richard Arkwright [wig maker & inventor] pioneered the water-propelled spinning frame which led to the mass production [automation] of cotton. This was truly revolutionary. The cotton manufacturers created a whole new class of people - the urban proletariat. The structure of society itself would never be the same*."

A. N. Wilson on Society

Chapter 1
Introduction

The aim of this PhD is to develop and evaluate automation methods to reduce the time and therefore the costs of conducting systematic reviews and other forms of evidence synthesis such as rapid reviews, meta-analyses and evidence overviews such as scoping searches. There are various definitions of what constitutes a systematic review, but in general a systematic review can be defined as aiming to identify and appraise all published and non-published evidence, using explicit and predefined methodological criteria to minimize bias and to synthesize and report the findings in a transparent manner that is open to criticism and correction from peers. There are several key steps to undertaking a systematic review (Figure 1) and much of the process is time intensive. Clinicians use systematic reviews to guide clinical decision making and they can also be used by policy makers to inform funding and policy decisions. In the hierarchy of evidence, systematic reviews are considered as the 'gold standard' for evaluating healthcare interventions. Scoping searches are often used to assess the size and scope of the research literature as a preliminary step to conducting a systematic review. Rapid reviews are a form of knowledge synthesis where some components of the systematic review process are simplified or omitted entirely in order to expedite the production of information[1].

Since the 18th century automation technologies have increasingly been applied to manufacturing to increase production[2]. In modern society, automation has been applied to various business sectors such as the automobile industry (robotic welding), air travel (fly-by-wire), retail industry (bar code scanners), and restaurants (touch screen ordering & conveyer-belt table service). Ideas that were once confined to the realm of science fiction are now realised such as automated postal delivery with drones[3], and driverless cars controlled by satellite navigation[4]. Healthcare is also benefiting from automation with IBM developing an artificial intelligence supercomputer for detecting lung cancer[5,6]. In contrast, the production of systematic reviews continues to rely considerably on human input and has not seen the same progress. Organisations are continuing to rely upon inefficient working practices that requires considerable human input.

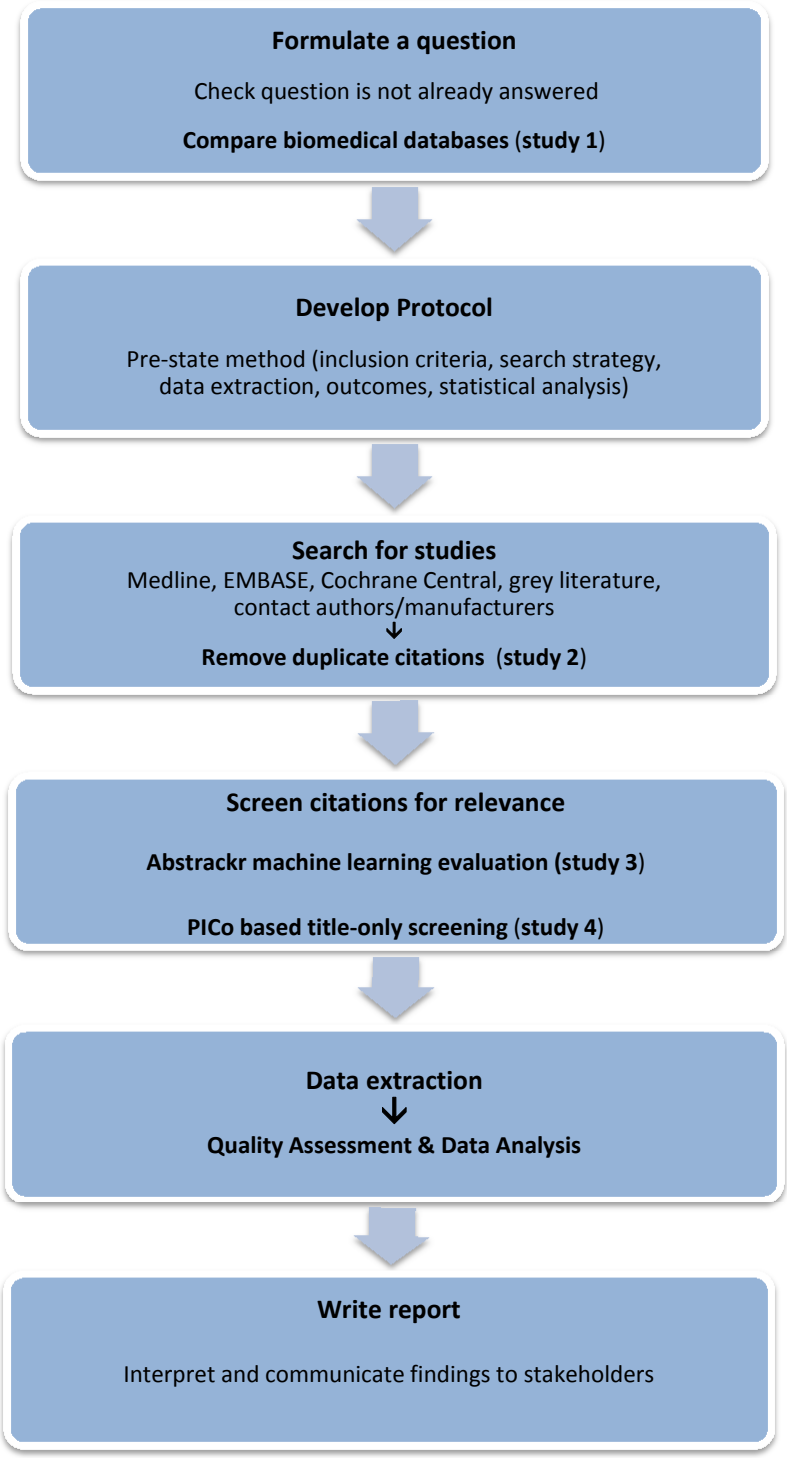


Figure 1. Key steps for conducting a systematic review and where studies for this PhD are focussed

1.1 History and development of systematic review

Prior to the development of systematic reviews the practice of healthcare was mostly opinion-based rather than evidence based. Criticism of this situation was first made by Archie Cochrane who, in 1971, wrote a report evaluating the UK National Health Service 'Effectiveness and Efficiency: Random Reflections on Health Services'[7]. In it he expressed concern that medicine should be based upon scientific evidence rather than the expert opinion of clinicians. His solution was to propose that medicine should be organised by specialty and sub-speciality and produce critical summaries that are adapted periodically of all relevant randomised controlled trials[7]. However, it took over twenty years before Archie Cochrane's vision commenced with the opening in 1992 of the United Kingdom Cochrane centre in Oxford. It has responsibility for the other UK and Ireland Cochrane groups, and for collaboration with different healthcare stake holders including the UK Department of Health, UK National Institute for Health Research and Oxford University. It was the first of 53 Cochrane groups to be founded within the international not for profit collaboration that now has contributors and centres from more than 100 countries.

1.2 Advantages of systematic reviews

Systematic reviews have the potential to detect the benefits and harms of healthcare interventions that otherwise might go undetected among a group of apparently conflicting individual trials. Trials cited selectively can contradict other studies and relying on just a subset of evidence from highly cited studies can inaccurately estimate the benefits and harms of treatment[8]. Similar trials are analysed together using statistical methods, known as meta-analysis, which incorporate the results of all studies into a single meta-calculation and thus increasing the power to detect the benefits or harms of healthcare interventions[9]. Such is the importance of systematic reviews that funding bodies are increasingly insisting that a systematic review must be performed as part of a clinical trial application to determine whether research gaps exist to justify funding further research[10]. Systematic reviews have the potential to save lives and resources. The earliest example, in healthcare, was a systematic review examining the effects of post-operative irradiation in women with early-stage breast cancer[11]. The 5 studies pooled together found that 5-year survival rates worsened with irradiation. The

effects of bad healthcare advice when applied to large populations can be devastating. In the 1950s Dr Spock, an American paediatrician, sold over 50 million copies in 42 countries of his best-selling book '*Baby and Child Care'* in which he advised parents to place infants face-down in their cots. The advice was wrong and *increased* the risk of sudden infant cot death syndrome. Yet, evidence that positioning infants face down increases the risk of death was available in the 1970s. If the evidence had been assessed systematically at the time these risks might have been addressed sooner, but because no systematic review was available the risks went unrecognised and as a consequence an estimated 60,000 children died[12]. Also, differentiating between effective and ineffective interventions ensures that finite resources are not wasted funding ineffective drugs and surgical procedures[13]. Systematic reviews also assess the quality of the evidence, such as estimating whether the results are biased due to methodological weaknesses, and assessing the strength of the evidence by calculating how important the findings are in terms of benefits to patients, e.g. a reduction in risk of death compared with current practice[14]. Also, because all the available evidence has been compiled together, systematic reviews are useful in indicating if there are knowledge gaps that require further research[15].

Systematic reviews are also used to provide evidence outside of medicine, such as determining the effects of different social policies[16], best educational practices[17] and the effects of different custodial sentences[18]. Clinicians and policy makers often require a rapid appraisal of the clinical evidence to inform policy decisions, often due to political urgency. However, this need for rapid appraisal is at odds with the time needed to produce governmental commissioned Health Technology Assessments (HTA). These consist of a systematic review and economic cost-effectiveness evaluation and typically take 6 to 12 months to complete[14,19]. Not surprisingly, Cochrane systematic reviews are more protracted due in part to their reliance on academic volunteers and on average, require 23 months to complete[20]. In contrast with such urgency, systematic reviewing has steadily become more time-consuming due to the introduction of additional reporting steps to improve transparency such as (1) the Summary of Findings tables which provides key information concerning the quality of evidence and the magnitude of effect of the interventions, (2) the Risk of Bias assessment which is a 6 domain assessment of methodological biases, and (3)

the 27-item Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist which is a minimum set of items for reporting in systematic reviews and meta-analyses.

Some systematic reviews now incorporate more complex indirect treatment comparisons using network meta-analysis[21] which are more time-consuming. This enables networks of trials, which may not have been directly compared against each other, to be evaluated in the context of a network of inferences. Other complex and time-consuming strategies include the analysis of clinical study reports that are held by regulatory agencies[22]. Incorporating such data can hamper evaluation due to the practical, and administrative difficulties in obtaining these reports. Additionally, the reports are often provided as an image-based file which prevents use of free text searches and thus hampers data extraction. These developments have helped to improve the reliability of systematic reviews, but consequently have made the task of systematic reviewing more complex and time-consuming.

1.3 Research growth

The growth in published research[23] has also been followed by the growth in published meta-analyses. For example in 1995, 429 meta-analyses were published in PubMed, however, by 2011 that figure had increased to 4739[24]. The scale of this problem facing the systematic review community is illustrated in Figure 2 which estimates the number of trials published from 1950 to 2010[23], and Figure 3 which estimates the number of systematic reviews published from 1990 to 2014[25].

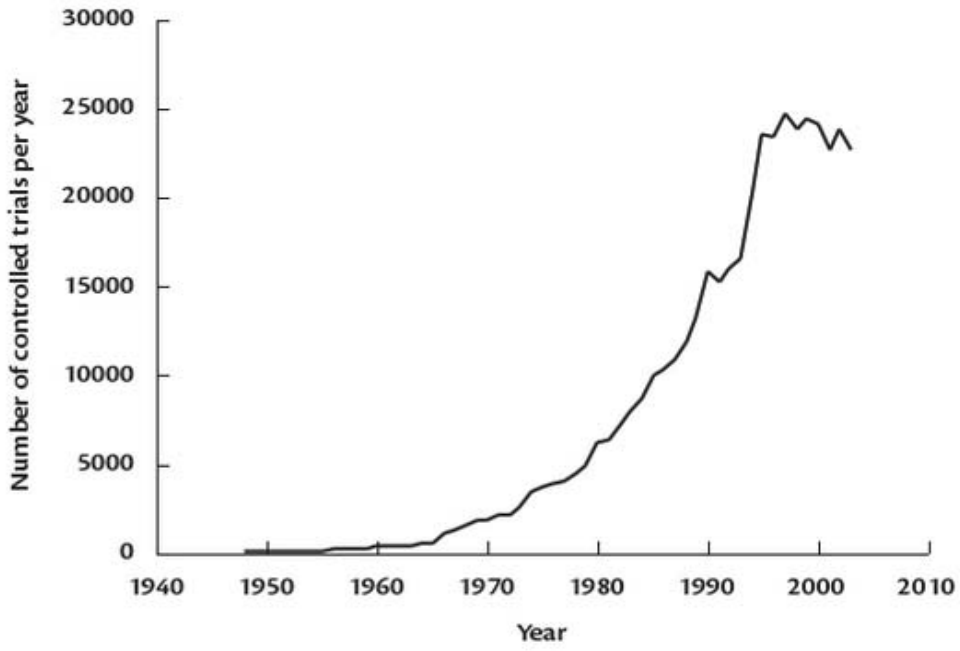


Figure 2. The estimated number of published trials from 1950 to 2010.
Source: Glasziou (2010) Evidence-Based Practice Workbook

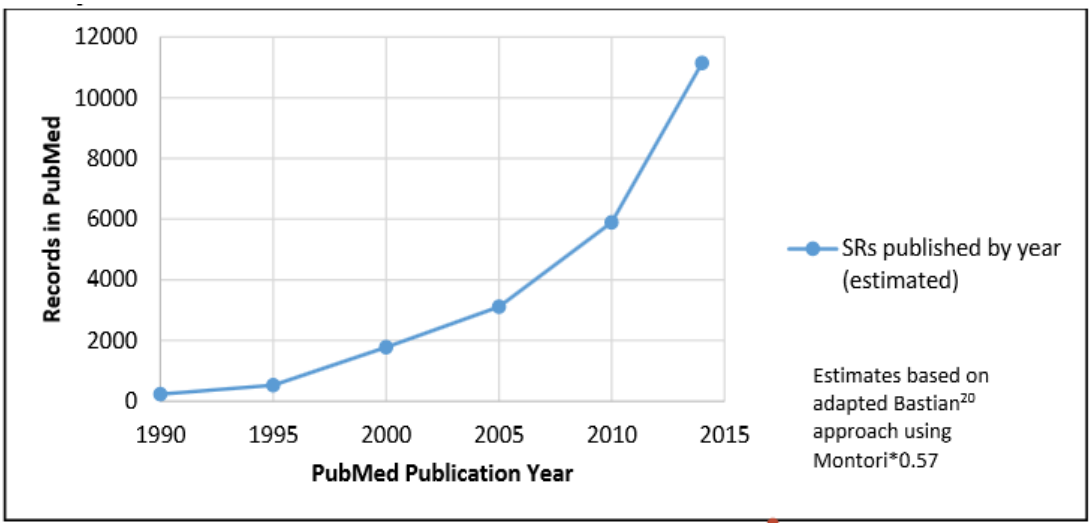


Figure 3. The estimated number of systematic reviews published from 1990 to 2014.
Source: Kleijen (2017) Evaluation of NIHR investment in Cochrane infrastructure and systematic reviews.

There are many barriers that researchers encounter whilst conducting a systematic review such as the phenomenon of multiple publishing of the same study data or 'Salami slicing' as it has been termed[26,27]. The consequence of multiple publications

for reviewers is additional time needed to sift through multiple study reports to ensure that all belong to the same study and are not mistakenly counted as a new study. Misidentifying multiple publications can lead to biased results and over-estimate the treatment effect[28]. The problem faced by reviewers when encountering 'salami science' can be considerable because subgroups are often reported so that numbers no longer match to the original study report and author names are re-ordered or changed. For example, in one Cochrane review the authors uncovered over 140 separate reports relating to a single trial of olanzapine[29]. Often, in such circumstances the only possible means to ascertain provenance is to contact the authors for clarification.

1.4 Updating systematic reviews

Systematic reviews can quickly become out of date when newly published research emerges. To maintain relevance the Cochrane Collaboration used to advise that systematic reviews should either be updated within two years of the first published version, or the previous update[14]. However, it has been shown that only 20% of Cochrane reviews are updated within two years after publication[30] leaving review groups struggling to remain relevant. There are several reasons for this, including availability of reviewers to commit time (much of the work is voluntary) and lack of financial resources. As a consequence, the original Cochrane policy of updating reviews regularly has been replaced with a policy based on prioritisation[31]. Non-Cochrane systematic reviews are also affected by the increased methodological complexity, and are at risk of being out of date by the time of publication[32]. Also, non-Cochrane reviews represent the majority of systematic reviews published, and have the most to gain from automation technologies (Figure 4)[25]. Additional barriers that researchers have cited for not conducting or updating systematic reviews include lack of reviewer motivation, limited academic credit and limited publishing formats[33]. Therefore, there is an urgency to improve current research practices by developing better methods and applications to assist reviewers.

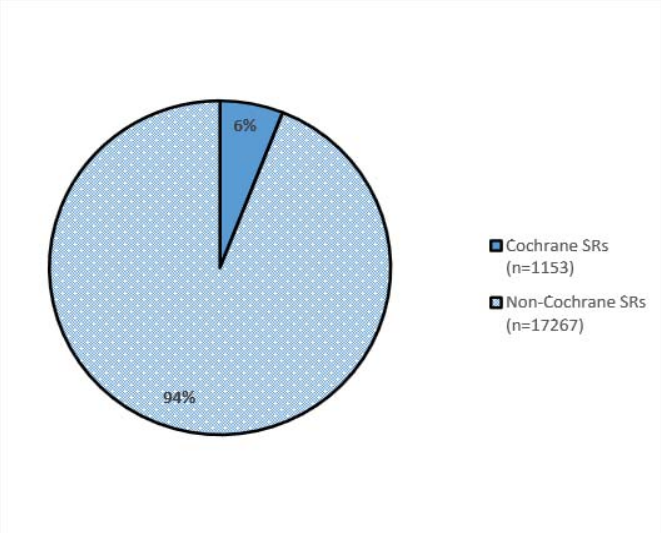


Figure 4 Percentage of all systematic reviews produced by Cochrane and other producers (total = 18,420; 2010-2015).

Source: Kleijen (2017) Evaluation of NIHR investment in Cochrane infrastructure and systematic reviews.

1.5 Current automation tools applied to systematic reviewing

The development of automation tools has already begun, with reviewing software aimed at assisting systematic reviewers such as ExaCT[34] that enable the reviewer to highlight relevant text such as the core components of a trial known as the 'PICOS' criteria (population, intervention, control, outcome, study design) which is the first step to ensure that the trial's study design matches the inclusion criteria of the systematic review. This tool is a useful aid to spot key information more quickly, but is unable to extract and input the information into systematic reviewing software such as RevMan. RobotReviewer is a program that automatically assesses risk of bias in clinical trials[35], and highlights the relevant text and extracts the information. It is an improvement on existing automated data extraction methods such as ExaCT, but is currently unable to equal the accuracy of a human screener. Web-based screening tools are now available such as Covidence[36], DistillerSR[37] and Rayyan[38] that visually enhance title and abstract screening by providing a user-friendly interface enabling the highlighting of subject specific keywords, automatically cross-checking screening decisions between co-reviewers, and ranking records according to the probability of relevance. Ranking records assists with identifying relevant

studies earlier so that hard-copies can be acquired sooner for data extraction to reduce delays.

These tools are under constant development and new features are being added based on feedback from users. Users have found many of the functions to greatly enhance collaborative review work by automatically cross-checking screening decisions, whilst other feature such as ranking trials by relevance have received mixed responses from reviewers with some users finding the feature unreliable. Such mixed reactions may be due to the complexity of different reviews and a thorough evaluation is needed to determine the accuracy of ranking records. Also, there are several automation tools offering similar functions and comparative evaluation is needed to determine the strengths and limitations of each product. Also, the development of toolkits[39] which have been designed to adapt HTA reports from one context or country to another may have an emerging role in automating systematic reviews. For example, a toolkit enables users to decide if new work is required or if existing HTA reports on the same or similar topics can be adapted for their purposes by prompting a series of questions relating to the quality and relevance of existing reports using a variety of domains such as safety, effectiveness, and cost-effectiveness. There are many areas within systematic reviews that are inefficient and would benefit from automation and some of these areas may rely on semi-automation such as processes that require the input of human operators either to enable the machine to learn from prior decisions, or because the workflow of data may be hampered by incompatibilities. For example, transferring information contained in a relational database to a spreadsheet requires an operator to edit the data to enable the information to be recognised and displayed correctly. Some automation tasks will be less hampered by technical problems and more suited to fully automated processes, such as with statistical analysis and generation of forest plots.

Expediting tasks that are time intensive for researchers, and therefore barriers to ensuring research remains relevant by the time of publication, are being explored with novel methods such as crowdsourcing. This involves engaging with large numbers of volunteers who each contribute to a project that would otherwise have required a full-time research team to achieve. For example, the cataloguing and

coding of bibliographic database records in PubMed and EMBASE is too basic to enable precise identification of studies during biomedical database searching. Nonetheless, this problem can be overcome with volunteers brought together via the internet to enrich bibliographic records with additional coding (e.g. PICOS coding for types of Participants, Intervention, Comparator, Outcomes and Study design). One such project has already begun with the Cochrane Dementia Group[40], whereby carers of people with dementia developed a specialist study-based register.

Significant savings can be made by maximising retrieval of relevant records since searches closely match the inclusion and exclusion criteria, eliminate duplicate records, and avoid unwittingly re-screening the studies during an update review. Also, a study-based specialised register solves the problem of multiple publications or 'salami science' and the resulting confusion and extra work this creates since records are electronically linked to the original paper, and avoids creating extra work downstream for the reviewers. For example, a study by the Cochrane Renal Group authors reported that in one review 56 reports were identified for just 14 trials[41], and they estimated that tracking down and linking these further publications added at least an extra four months to completing their review. Without a study based register these problems are repeated for each new review title and its subsequent updates. Also, with a specialist register coding is only performed once for the original trial not for further reports. The current practice of updating a Cochrane review every two years ideally should be replaced with instant meta-analysis whereby meta-analysis is performed whenever a new trial is published. To achieve such an efficient model of production the current inefficiencies need to be addressed by adopting or developing methods from computer science. These problems have also recently been recognised in a (2017) UK Department of Health report[25] which recommends that the Cochrane Collaboration "should work on developing expertise and processes to get better and quicker at producing reviews".

1.6 Summary

Systematic reviewing has developed considerably both methodologically and organisationally since its inception in the 1970s. Its importance to the advancement of health research in recognised by both the research community and funding

bodies. The growth in research has not been met by a growth in automation technologies and systematic review teams are unable to keep pace with the data influx. Better working practices are urgently needed that incorporate automation technologies to ensure that research findings remain relevant. Considerable gaps exist in the development and evaluation of automation technologies and to that aim this PhD is focussed on addressing those gaps in our current knowledge.

Chapter 2
Research proposal

The research aims of this PhD are to develop strategies to expedite the production of systematic reviews by developing and evaluating new methods to overcome the inefficiencies with current practices. Subsequently, four research projects (as shown in Figure 1) were conceived. The rationale for each research study was predicated upon several factors including personal knowledge of existing 'bottle-necks' that affect reviewers but which have not been previously investigated, or have not been fully developed. From this study 1 was conceived (Comparison of Biomedical Databases) which provided new insights into the scope and reliability of databases. Study 2 (Deduplication) was conceived, in part, due to the lack of progress shown by commercial software companies to advance current practices and therefore new methods were needed to overcome current limitations with deduplication. Study 3 (Abstrackr predictive screening) was conceived to investigate the reliability of emerging technologies that are being applied to systematic reviews and that have the potential to save resources on tasks that are time-intensive. Study 4 (PICo based title-only citation screening) was conceived to develop new strategies to advance semi-automation by expediting screening using strategies that could either replace or complement current automation technologies.

2.1 A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension

Aim

The aim of the study was to compare major bibliographic databases to determine which databases are best for identifying systematic reviews and how many databases need to be searched to identify all relevant records. The rationale for this study was the research gap in our understanding of how well databases performed specifically to identify systematic reviews and how many databases are needed to identify all information, including the reliability and comprehensiveness.

Objectives:

1. Identify major bibliographic databases (MEDLINE, EMBASE, Cochrane library, PubMed-Health, DARE, Epistemonikos, TRIP) that index systematic reviews.

2. Develop search strategies for each database to identify systematic reviews of intervention studies for the treatment of hypertension.

3. Screen each database for relevant and irrelevant systematic reviews and compare screening decisions between relational database to ensure consistency with screening decisions.

4. Evaluate the performance of each database to identify relevant studies using sensitivity and precision. Using venn diagrams determine the number of databases needed to be searched to identify all relevant systematic reviews that were identified as relevant, i.e. the reference set.

2.2 Better duplicate detection for systematic reviewers: evaluation of systematic Review Assistant-Deduplication Module

Aim

The aim of the study was to develop methods to identify and remove duplicate records retrieved from biomedical database searches. The rationale for this study was predicated upon the poor performance of current practices. Existing methods of duplicate detection are unsatisfactory due to the poor performance, and therefore a new approach was needed to expedite duplicate detection and reduce workload by identifying duplicate citations with greater accuracy than the current method available in EndNote™ and similar bibliographic reference management software.

Objectives:

1) Evaluate the accuracy of the default auto-deduplication in EndNote™ against the benchmark.

2) Evaluate the accuracy of the new deduplication algorithm against the benchmark.

3) Compare the accuracy of the new algorithm against the performance of EndNote™ and calculate the sensitivity and specificity.

4) Identify why records were wrongly classified (by algorithm), i.e. false positive and false negative, and incorporate the findings into each iteration of the algorithm.

5) Validate the accuracy of the final optimised algorithm using a series of different datasets from intervention studies and screening tests, using different topic specialities (cytology screening, stroke, and haematology).

2.3 Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers

Aim

The aim of the study was to evaluate the feasibility of using semi-automated screening methodologies to expedite title and abstract screening. The study was developed to address one of the current research gaps of limited validation of automation methods.

Objectives:

1. Evaluate the recall accuracy of a semi-automated, machine learning citation screening program - Abstrackr.

2. Evaluate the workload saving of a semi-automated machine learning citation screening program - Abstrackr.

3. Perform sensitivity analyses on datasets with indistinct groups of participants.

2.4 PICo based title-only screening to expedite reviewing

Aim

The aim of this study was to develop and evaluate a new method to expedite the screening of study citations retrieved from biomedical database searches using a PICo based title-only screening method. The rationale for this study was the current unsatisfactory development with semi-automated citation screening applications which prompted the need to explore and develop an alternative method to reduce the workload with citation screening.

Objectives:

1. Survey the literature to obtain a sample of different datasets from previously published systematic reviews to evaluate the screening methodology.

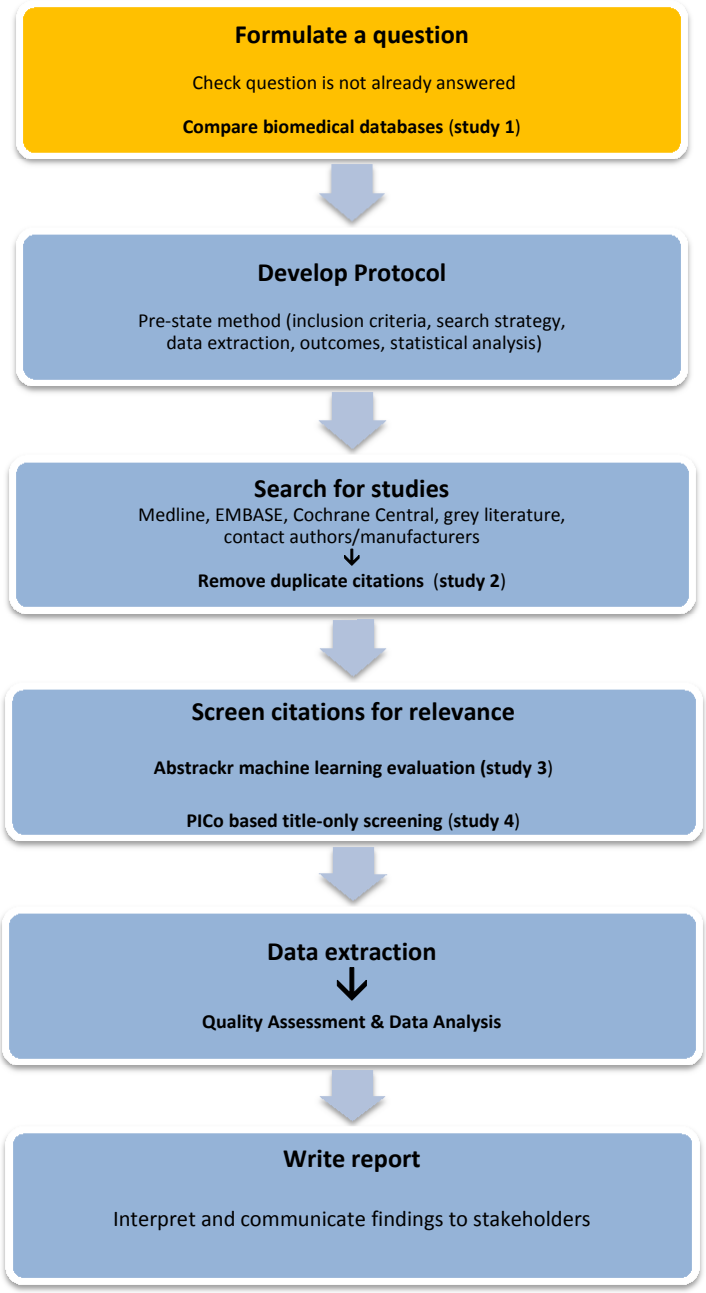
2. Develop keyword searching criteria and Boolean operators suitable for restrictive title filed only searching.

3. Generate a list of search terms from the systematic reviews inclusion criteria based upon the PICo criteria. Generate a list of synonyms for the PICo terms.

4. Evaluate the screening results for recall against the included studies. Evaluate the percentage reduction in screening effort.

Chapter 3

Biomedical database coverage



Key steps for conducting a systematic review and where studies for this PhD are focussed

In Chapter 1, the problem of the growth in published systematic reviews was highlighted. Prior to conducting a systematic review, it is necessary to determine whether the research question has been previously answered in an existing review, and thus avoiding wasteful replication of research. Answering this question requires a search of bibliographic databases, but this can be time-consuming because it is not known which biomedical database is the best to search.

This question led to the development of a research study comparing seven biomedical databases to determine how many databases are necessary to identify all relevant systematic reviews on a given topic. This was the first published study that attempted to answer this question, as literature searches did not find equivalent or similar research. The research methods, data collection and analysis are described in this chapter and the findings discussed.

Summary

The published study demonstrated that no single database could identify all published systematic reviews on the topic of hypertension. This is due to the content and coverage of each database, but also the limitation of using systematic review search filters which can reduce the sensitivity of the search results. Nonetheless for this topic, EMBASE was the best performing database when assessed by sensitivity. The Cochrane library, which also indexes systematic reviews from other databases such as DARE and HTA, was the best performing when assessed according to specificity. Regardless of which database is chosen, there is a trade-off between sensitivity and specificity and researchers need to choose which database is most appropriate. However, given the Cochrane library had the best specificity and was the second best performing database for sensitivity, researchers may be more likely to use this as the default option, before widening the search to include other databases.

3.1 A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension

Systematic Reviews (2016) **5**:27. doi: 10.1186/s13643-016-0197-5.

John Rathbone

Matt Carter

Tammy Hoffmann

Paul Glasziou

Abstract

Background

Bibliographic databases are the primary resource for identifying systematic reviews of healthcare interventions. Reliable retrieval of systematic reviews depends on the scope of indexing used by database providers. Therefore, searching one database may be insufficient, but it is unclear how many need to be searched. We sought to evaluate the performance of seven major bibliographic databases for the identification of systematic reviews for hypertension.

Methods

We searched seven databases (Cochrane library, Database of Abstracts of Reviews of Effects (DARE), Excerpta Medica Database (EMBASE), Epistemonikos, Medical Literature Analysis and Retrieval System Online (MEDLINE), PubMed Health and Turning Research Into Practice (TRIP)) from 2003 to 2015 for systematic reviews of any intervention for hypertension. Citations retrieved were screened for relevance, coded and checked for screening consistency using a fuzzy text matching query. The performance of each database was assessed by calculating its sensitivity, precision, the number of missed reviews and the number of unique records retrieved.

Results

Four hundred systematic reviews were identified for inclusion from 11,381 citations retrieved from seven databases. No single database identified all the retrieved systematic reviews for hypertension. EMBASE identified the most reviews (sensitivity 69 %) but also retrieved the most irrelevant citations with 7.2 % precision (Pr). The sensitivity of the Cochrane library was 60 %, DARE 57 %, MEDLINE 57 %, PubMed Health 53 %, Epistemonikos 49 % and TRIP 33 %. EMBASE contained the highest number of unique records ($n = 43$). The Cochrane library identified seven unique records and had the highest precision (Pr = 30 %), followed by Epistemonikos ($n = 2$, Pr = 19 %). No unique records were found in PubMed Health (Pr = 24 %) DARE (Pr = 21 %), TRIP (Pr = 10 %) or MEDLINE (Pr = 10 %). Searching EMBASE and the Cochrane library identified 88 % of all systematic

reviews in the reference set, and searching the freely available databases (Cochrane, Epistemonikos, MEDLINE) identified 83 % of all the reviews.

The databases were re-analysed after systematic reviews of non-conventional interventions (e.g. yoga, acupuncture, exercise) were removed. Similarly, no database identified all the retrieved systematic reviews. EMBASE identified the most relevant systematic reviews (sensitivity 73 %) but also retrieved the most irrelevant citations with Pr = 5 %. The sensitivity of the Cochrane database was 62 %, followed by MEDLINE (60 %), DARE (55 %), PubMed Health (54 %), Epistemonikos (50 %) and TRIP (31 %). The precision of the Cochrane library was the highest (20 %), followed by PubMed Health (Pr = 16 %), DARE (Pr = 13 %), Epistemonikos (Pr = 12 %), MEDLINE (Pr = 6 %), TRIP (Pr = 6 %) and EMBASE (Pr = 5 %). EMBASE contained the most unique records (n = 34). The Cochrane library identified seven unique records. The other databases held no unique records.

Conclusions

The coverage of bibliographic databases varies considerably due to differences in their scope and content. Researchers wishing to identify systematic reviews should not rely on one database but search multiple databases.

Introduction

Systematic reviews provide the best evidence of the effects of healthcare interventions [1]. However, identifying systematic reviews can be time-consuming and haphazard because no database covers all health topics [2]. Therefore, searching several databases is a necessity when seeking health research, including systematic reviews. With the growth [3] and scatter of research [4], finding relevant and up-to-date information is becoming increasingly difficult. Moreover, clinicians who perform quick clinical queries with one database often lack the training and skills to run efficient searches and subsequently produce imprecise results [5]. Understandably, there is currently no specific guidance on which databases should be searched to find systematic reviews, only general advice to search widely. For example, researchers planning a systematic review are recommended to first search for existing reviews which answer the research question to avoid duplicating research [6], but it is unclear which is the best database to search or how many should be searched.

The aim of this study was to evaluate seven databases—the Cochrane library, the Database of Abstracts of Reviews of Effects (DARE), Excerpta Medica Database (EMBASE), Epistemonikos, Medical Literature Analysis and Retrieval System Online (MEDLINE), PubMed Health and Turning Research Into Practice (TRIP) to determine their coverage of systematic reviews assessing effectiveness of interventions of a typical high-prevalence condition, hypertension, and to determine how many databases require searching to identify all relevant systematic reviews.

Methods

We searched seven databases (EMBASE, MEDLINE, the Cochrane library (inc. CDSR, DARE and HTA), Epistemonikos, PubMed Health, DARE and TRIP) for systematic reviews of any treatment interventions for hypertension from 2003 to Jan 2015 (see Fig. 1). The databases were chosen because of their prominence as research databases that index systematic reviews. We used an open definition of systematic review which included reviews stated or described as being a systematic review or meta-analysis. Reports and summaries of evidence were excluded. PICO criteria were defined as follows: participants, i.e. people with hypertension by any

definition; interventions, any; comparator, any; and outcomes, change in blood pressure. Systematic review filterers incorporated into the databases were selected to increase search sensitivity. For MEDLINE, we used the Montori filter [7]. Citations retrieved were imported into separate EndNote™ X7 libraries, and then titles and abstracts were screened for relevance by one reviewer. Reviews of pre-hypertension, ophthalmic, pulmonary, pregnancy-related hypertension or hepatic hypertension were excluded. Greasemonkey scripts were used to assist with the retrieval of the contents of web pages which did not have full citation download options (see supplementary appendix A).

PubMed Health

(hypertension OR antihypertensive) NOT (pregnancy OR intraocular OR interocular OR ophthalmic)
Limited to Article Type – Systematic Reviews
Custom range date filter was not working so no date range applied

Epistemonikos

(title:(hypertension OR antihypertensive) OR abstract:(hypertension OR antihypertensive))
[Filters: min_year=2003, max_year=2015, classification=systematic-review]

Trip

(title:hypertension or antihypertensive) (not pregnancy) from:2003 to:2015

DARE

1 MeSH DESCRIPTOR hypertension IN DARE
2 MeSH DESCRIPTOR Antihypertensive Agents IN DARE
3 (hypertens*) OR (antihypertens* or anti-hypertens*) IN DARE
4 #1 OR #2 OR #3
5 (pregnancy OR ophthalmic OR interocular OR intraocular) IN DARE
6 #4 NOT #5
7 * IN DARE FROM 2003 TO 2015
8 #6 AND #7

Medline (Ovid)

1 *Hypertension/ (138456)
2 exp *Antihypertensive Agents/ (119724)
3 (hypertens* or antihypertens* or anti-hypertens*).tw. (315155)
4 or/1-3 (406228)
5 (pregnancy or ophthalmic or interocular or intraocular).tw. (332321)
6 4 not 5 (387506)
7 (medline or systematic review).tw. or meta-analysis.pt. (110858)
8 6 and 7 (3108)
9 limit 8 to yr="2003 -Current" (2419)

Embase (Elsevier)

#5 #4 AND [2003-2015]/py 3,829
#4 #3 AND ([cochrane review]/lim OR [systematic review]/lim OR [meta analysis]/lim) 4,762
#3 #1 NOT #2 436,107
#2 pregnan*:ab,ti OR ophthalm*:ab,ti OR interocular:ab,ti OR intraocular:ab,ti 623,580
#1 hypertens*:ab,ti OR antihypertens*:ab,ti OR 'anti-hypertensive':ab,ti OR 'anti-hypertensives':ab,ti 465,036

Cochrane Library
CDSR & HTA & HTA

#1 MeSH descriptor: [Hypertension] this term only
#2 MeSH descriptor: [Antihypertensive Agents] explode all trees
#3 hypertens* or antihypertens* or anti-hypertens*:ti,ab,kw (Word variations have been searched)
#4 #1 or #2 or #3
#5 pregnancy or ophthalmic or interocular or intraocular:ti,ab,kw (Word variations have been searched)
#6 #4 not #5 Publication Year from 2003 to 2015, in Cochrane Reviews (Reviews and Protocols) and Technology Assessments

Fig. 1 Search strategies

Citations were coded in EndNote™ X7 as either a systematic review or not. Screening decisions in one database were cross-checked against the other six databases to ensure consistency using a title-matching database query. The query incorporated a fuzzy text matching algorithm [8, 9] to account for differences with punctuation or syntax errors. Where screening decisions were found to be inconsistent, these were re-examined and standardised across the databases. Where databases (e.g. PubMed Health) used the Cochrane plain language title rather than the original full title, these were changed to the full title for consistency with other databases*.

Data analysis

The performance of each database was assessed by calculating the sensitivity (number of relevant studies/reference set × 100); the precision (number of relevant studies/number of studies retrieved × 100); the number missed (reference set − number of relevant studies); and the number of unique records, i.e. records only found in one database. The reference set is the total of unique systematic reviews identified across all the databases. Records identified as being unique were double-checked for accuracy using a title search within the (online) comparator bibliographic databases without the systematic review search filters applied. A secondary analysis was performed by removing all non-conventional treatments, i.e. systematic reviews that are not prescribed drugs, e.g. yoga, acupuncture, herbal medicine, and exercise programmes, from the databases and re-calculated to provide results reflecting the type of quick clinical queries clinicians would run.

---

*For additional methodological details see Supplementary appendix A

Results

There were 400 systematic reviews (the reference set) identified for inclusion from a total of 11,381 citations retrieved from seven databases. No database identified all 400 included systematic reviews of interventions for hypertension (Table 1). EMBASE retrieved the highest number of relevant reviews (*n* = 276) with a sensitivity (*s*) of 69.0 %, followed by Cochrane (*n* = 240, *s* = 60.0 %), DARE (*n* = 228, *s* = 57.0 %), MEDLINE (*n* = 228, *s* = 57.0 %), PubMed Health (*n* = 212, *s* = 53.0 %), Epistemonikos (*n* = 195, *s* = 48.8 %) and TRIP (*n* = 131, *s* = 32.8 %). EMBASE contained the largest number of unique records (*n* = 43) but had the lowest precision (Pr, 7.2 %). Cochrane contained seven unique records and had the highest precision (29.9 %), followed by Epistemonikos (*n* = 2, Pr = 19.2 %). No unique records were found in PubMed Health (Pr = 23.6 %), DARE (Pr = 20.8 %), TRIP (Pr = 9.7 %) or MEDLINE (Pr = 9.6 %). Searching the two databases with the highest sensitivity and unique records (EMBASE and the Cochrane library) identified 88 % of the reference set (Fig. 2). Searching the Cochrane library, MEDLINE and Epistemonikos identified 83 % of the reference set (Fig. 3).

Table 1 Performance of bibliographic databases identifying relevant systematic reviews of interventions for treating hypertension

Database	Reviews relevant	Reviews missed	Total citations	Sensitivity (s) %	Precision (Pr) %	Unique records
EMBASE	276	124	3836	69.0	7.2	43
Cochrane	240	160	802	60.0	29.9	7
DARE	228	172	1098	57.0	20.8	0
MEDLINE	228	172	2374	57.0	9.6	0
PubMed Health	212	188	899	53.0	23.6	0
Epistemonikos	195	205	1017	48.8	19.2	2
TRIP	131	269	1355	32.8	9.7	0

Reference set of included systematic review (*n* = 400)

Fig. 2 Proportion of reference set (*n* = 400) retrieved by searching *EMBASE* and the *Cochrane* library, resulting in the identification of 88 % (*n* = 352) of total reviews

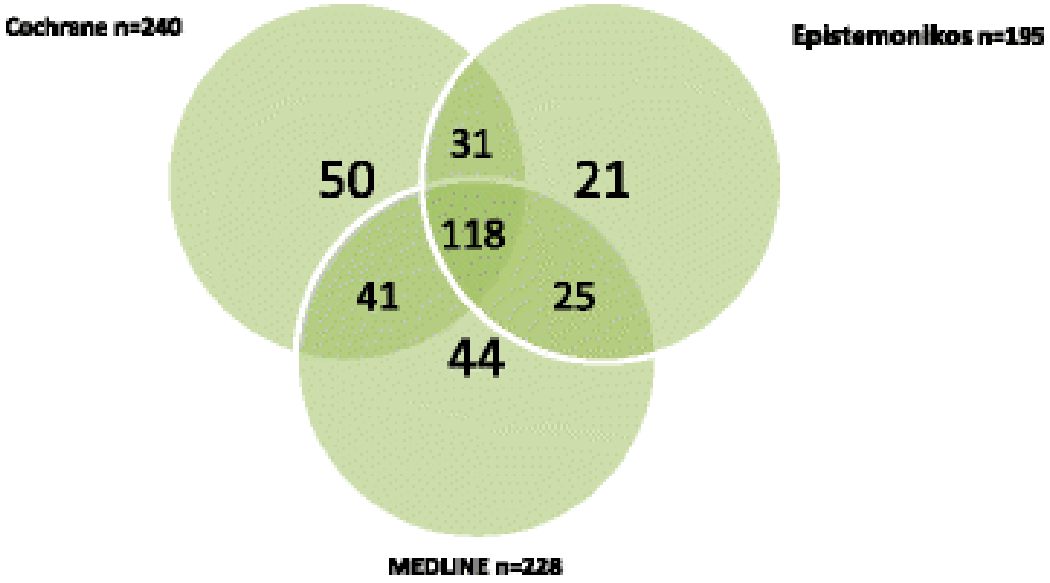


Fig. 3 Proportion of reference set (*n* = 400) retrieved by searching *Cochrane*, *Epistemonikos* and *MEDLINE*, resulting in the identification of 83 % (*n* = 330) of total reviews

After removing 168 non-conventional medical interventions for hypertension, e.g. yoga, acupuncture, herbal medicine, and exercise programmes, there were 232 systematic reviews remaining in the reference set. Again, no database identified all included systematic reviews of conventional interventions for hypertension (Table 2). EMBASE retrieved the highest number of relevant records (*n* = 169) with a

sensitivity of 72.8 %, followed by the Cochrane library (*n* = 143, *s* = 61.6 %), MEDLINE (*n* = 138, *s* = 59.5 %), DARE (*n* = 127, *s* = 54.7 %), PubMed Health (*n* = 126, *s* = 54.3 %), Epistemonikos (*n* = 116, *s* = 50.0 %) and TRIP (*n* = 72, *s* = 31.0 %). EMBASE contained the largest number of unique records (*n* = 34) but had the lowest precision (Pr = 4.5 %). Cochrane contained seven unique records and had the highest precision (Pr = 20.3 %). No unique records were found in PubMed Health (Pr = 15.5 %), DARE (Pr = 12.7 %), Epistemonikos (Pr = 12.4 %), MEDLINE (Pr = 6.0 %) or TRIP (Pr = 5.5 %).

Table 2 Performance of bibliographic databases identifying relevant systematic reviews of interventions for treating hypertension (excluding non-conventional treatments)

Database	Reviews relevant	Reviews missed	Total citations	Sensitivity (s) %	Precision (Pr) %	Unique records
EMBASE	169	63	3722	72.8	4.5	34
Cochrane	143	89	704	61.6	20.3	7
MEDLINE	138	94	2282	59.5	6.0	0
DARE	127	105	998	54.7	12.7	0
PubMed Health	126	106	812	54.3	15.5	0
Epistemonikos	116	116	938	50.0	12.4	0
TRIP	72	160	1320	31.0	5.5	0

Reference set of included systematic review (*n* = 232)

Discussion

Seven databases were searched—the Cochrane library, DARE, EMBASE, Epistemonikos, MEDLINE, PubMed Health and TRIP—to determine their coverage of systematic reviews of interventions for hypertension. No single database retrieved the entire reference set of 400 reviews; EMBASE had the highest sensitivity of 69 % but would still miss 124 reviews. Searching both the Cochrane library and EMBASE identified 88 % of the reference set. EMBASE, however, is a subscription service and many institutions do not subscribe to EMBASE, which may limit some clinicians from performing clinical queries. Nevertheless, in the example used in this study, searching the Cochrane library, MEDLINE and Epistemonikos retrieves 83 % of the reference set.

Our findings have illustrated that despite the broad scope of many bibliographic databases, relying on one or two to identify a systematic review is not always possible, and wider search should be considered to ensure systematic reviews are not missed.

Strengths and limitations

We used systematic review filters to increase precision during the search for hypertension reviews, which can reduce the sensitivity. Therefore, records classed as unique were cross-checked with the comparator databases by searching in title fields without applying the filter to ensure the record was genuinely unique rather than missed due to filtering. However, this procedure was not performed where systematic reviews were found in two or more databases, and therefore, some reviews may have been missed due to use of filters or the reviews being inadequately coded in the databases. However, reduced sensitivity will have affected all databases since filters were applied universally. Discarding search filters, however, is impractical due to the large number of records that would be retrieved. Screening was performed by one reviewer with the potential for screening errors between databases; therefore, to ensure screening decisions were consistent, a fuzzy text matching query [10] was used. Our case study did not include every bibliographic database available, but we included seven major databases, including the two largest (EMBASE and MEDLINE); however, the results may not be

applicable to specialist databases if they are not indexed in MEDLINE, EMBASE or the Cochrane library. Our focus was limited to one clinical condition (hypertension), but other clinical topics are also likely to be dispersed throughout these databases without a single database containing all records. Other study designs such as prognostic and diagnostic studies were not evaluated, and database searches for this type of study design may perform differently. The DARE database provided a search platform with good overall sensitivity and precision, but funding for DARE ceased at the end of March 2015 ([11]), and as it is no longer being updated, this database will increasingly become less sensitive for identifying systematic reviews.

Conclusions

This case study demonstrated that relying on a single database is insufficient to identify all relevant systematic reviews. Depending on the database used, the chances of finding the proportion of relevant reviews ranged from 33 to 69 %, and therefore, searching should not be restricted to two major databases; instead, a search of additional databases should be performed to determine if a review title exists. Further research is warranted to assess how these findings might extend to other topic areas and study designs.

Abbreviations

CDSR: Cochrane database of systematic reviews

DARE: Database of abstracts of reviews of effects

EMBASE: Excerpta medica database

HTA: Health technology appraisal

MEDLINE: Medical literature analysis and retrieval system online

PICO: Participants, interventions, comparators, outcomes

TRIP: Turning research into practice

Declarations

Acknowledgements

We thank Sarah Thorning for running the database searches.

Funding

NHMRC Australia Fellowship: GNT05275

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed to the study concept and design. JR devised the testing and analysis of the database. MC wrote the fuzzy text matching title code and devised the truth table. JR drafted the initial manuscript. TH, PG and MC contributed to the manuscript and all the revisions. All authors read and approved the final manuscript.

References

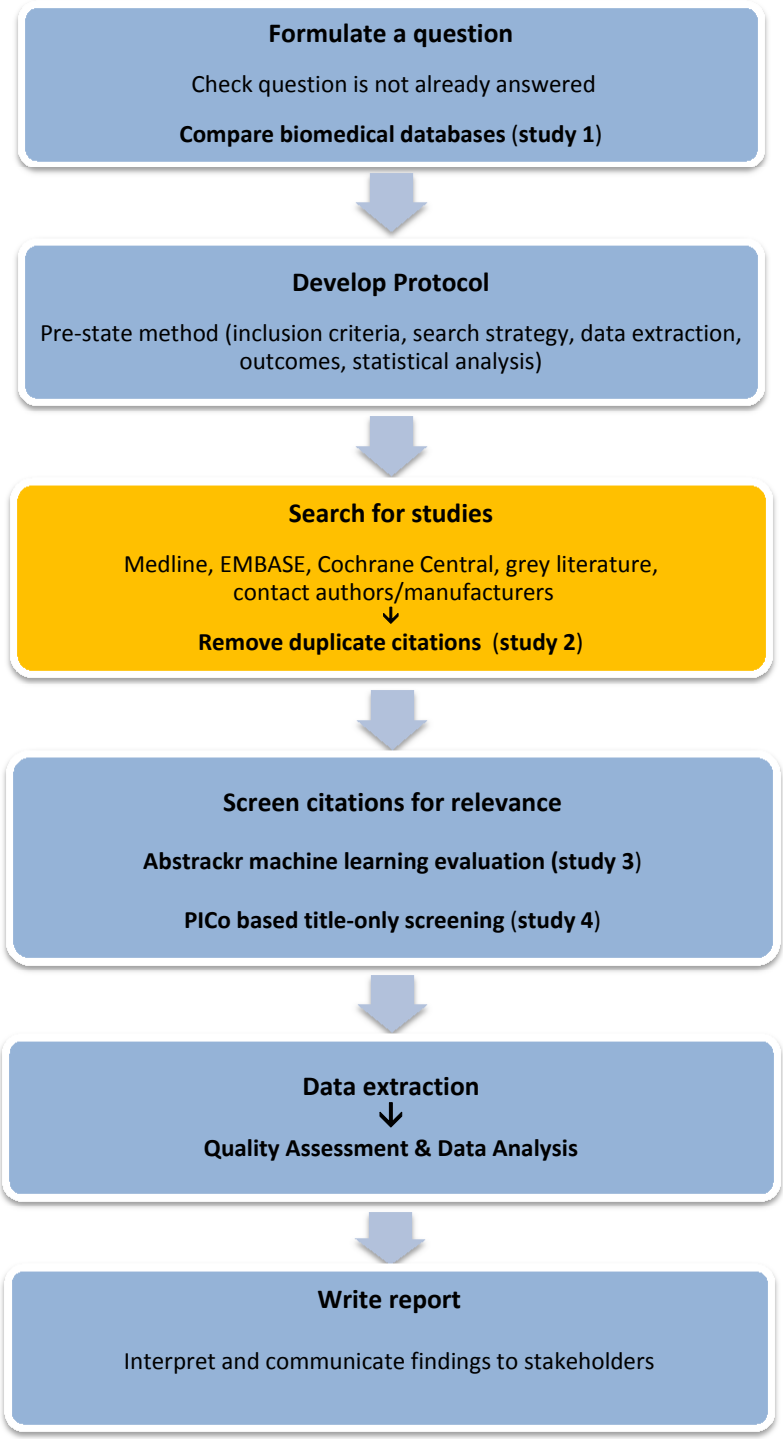
1. **Oxford Centre for Evidence-Based Medicine—levels of evidence** [*http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/*] Accessed 23-01-2015.
2. Shariff SZ, Bejaimal SA, Sontrop JM, Iansavichus AV, Haynes RB, Weir MAGA. **Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches**. J Med Internet Res. 2013;15:e164.*PubMed CentralView ArticlePubMedGoogle Scholar*
3. Bastian H, Glasziou P, Chalmers I. **Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?** PLoS Med. 2010;7:e1000326.*PubMed CentralView ArticlePubMedGoogle Scholar*
4. Hoffmann T, Erueti C, Thorning S, Glasziou P. **The scatter of research: cross sectional comparison of randomised trials and systematic reviews across specialties**. BMJ. 2012;344:e3223.*PubMed CentralView ArticlePubMedGoogle Scholar*
5. Ely J, Osheroff J, Ebell M, Chambliss M, Vinson D, Stevermer J, et al. **Obstacles to answering doctors' questions about patient care with evidence: qualitative study**. BMJ. 2002;324:710.*PubMed CentralView ArticlePubMedGoogle Scholar*
6. **Systematic reviews: CRD's guidance for undertaking reviews in health care** [*https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf*]. Accessed 22-09-2015.
7. Montori VM, Wilczynski NL, Morgan D, Haynes RB. **Optimal search strategies for retrieving systematic reviews from Medline: analytical survey.** BMJ. 2005;330(December):68.*PubMed CentralView ArticlePubMedGoogle Scholar*
8. **Fuzzy text matching algorithm** [*https://github.com/CREBP/CREBP-DB-Comparison*]. Accessed 24-09-2015.
9. Rathbone J, Carter M, Hoffmann T, Glasziou P**. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module.** Syst Rev. 2015;14;4:6.

10. **Systematic Review Assistant-Deduplication Module** [*http://crebp-sra.com*].

11. **Centre for Reviews and Dissemination**
[*http://www.crd.york.ac.uk/CRDWeb/NewsPage.asp*]. Accessed 23-01-2015.

Chapter 4

Duplicate detection within bibliographic records



Key steps for conducting a systematic review and where studies for this PhD are focussed

In Chapter 3, the problems of relying on a single database to identify systematic reviews was highlighted. Even searching several databases is no guarantee that all published reviews are identified, due in part to differences in the scope of each database provider, sensitivity of systematic review filters, technical errors with plain language summaries being used in place of the original title, and technical problems with databases. The most reliable database (EMBASE) was also the least precise, requiring thousands of records to be screened. Due to the overlapping content of different biomedical databases, duplicate records are inevitably retrieved during searches for both systematic reviews and randomised controlled trials (as highlighted in Chapter 1). From this problem arose the question: How can the identification and removal of duplicate records from bibliographic databases be improved?

In Chapter 4 the research undertaken to identify duplicate records more effectively is described. Four deduplication algorithms were developed and modified by incorporating the findings from each version. To test the performance of an algorithm, a 'gold standard' reference set of citations, from a published systematic review, was created by coding citations as either a unique or duplicate record. Each algorithm and the auto-deduplication facility in EndNote™ were tested against the reference set. Following this iterative process of testing and developing the algorithms, the best performing algorithm was selected for validation testing using datasets from three systematic reviews to determine if the initial findings were replicable. The implications for this new method of duplicate detection within the research community are discussed and recommendations are provided for additional research to further improve duplicate detection.

Summary

This study was published on 14th January 2015[42] and has subsequently been cited 15 times. The detection of duplicate records by the new algorithm was greater than EndNote™, achieving both higher sensitivity and specificity. The algorithm is being used at the Centre for Research in Evidence-Based Practice (CREBP), and following the publication of this study, the deduplication application has been made available to other academic researchers and information specialists through the CREBP web page. Interest in the algorithm from research groups is growing

following the study's publication. Links with the Melbourne Cochrane group have been established with the aim of collaborating with the development of automation tools, including incorporating the deduplication tool into their existing Covidence screening software[36]. The Cochrane Collaboration has also expressed interest as currently their deduplication tool is based upon the deduplication algorithm used in EndNote™. The deduplication research was further disseminated with a presentation at The European Public Health conference in Milan, Italy in 2015 and received interest from participants active in public health research. A revised second version is planned that will enable users to pre-select the degree of sensitivity and specificity according to their needs.

4.1 Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module

Systematic Reviews (2015) **4**:6. doi: 10.1186/2046-4053-4-6.

John Rathbone

Matt Carter

Tammy Hoffmann

Paul Glasziou

Abstract

Background

A major problem arising from searching across bibliographic databases is the retrieval of duplicate citations. Removing such duplicates is an essential task to ensure systematic reviewers do not waste time screening the same citation multiple times. Although reference management software use algorithms to remove duplicate records, this is only partially successful and necessitates removing the remaining duplicates manually. This time-consuming task leads to wasted resources. We sought to evaluate the effectiveness of a newly developed deduplication program against EndNote™.

Background

A literature search of 1,988 citations was manually inspected and duplicate citations identified and coded to create a benchmark dataset. The Systematic Review Assistant-Deduplication Module (SRA-DM) was iteratively developed and tested using the benchmark dataset and compared with EndNote's default one step auto-deduplication process matching on ('author', 'year', 'title'). The accuracy of deduplication was reported by calculating the sensitivity and specificity. Further validation tests, with three additional benchmarked literature searches comprising a total of 4,563 citations were performed to determine the reliability of the SRA-DM algorithm.

Results

The sensitivity (84%) and specificity (100%) of the SRA-DM was superior to EndNote™ (sensitivity 51%, specificity 99.83%). Validation testing on three additional biomedical literature searches demonstrated that SRA-DM consistently achieved higher sensitivity than EndNote™ (90% vs 63%), (84% vs 73%) and (84% vs 64%). Furthermore, the specificity of SRA-DM was 100%, whereas the specificity of EndNote™ was imperfect (average 99.75%) with some unique records wrongly assigned as duplicates. Overall, there was a 43% relative increase in the number of duplicates records detected with SRA-DM compared with EndNote™ auto-

deduplication.

Conclusions

The Systematic Review Assistant-Deduplication Module offers users a reliable program to remove duplicate records with greater sensitivity and specificity than EndNote™. This application will save researchers and information specialists time and avoid research waste. The deduplication program is freely available online.

Background

Identifying trials for systematic reviews is time consuming: the average retrieval from a PubMed search produces 17,284 citations [1]. The biomedical databases MEDLINE[2] and EMBASE[3] contain over 41 million records, and about one million records are added annually to EMBASE[3] (which now also includes MEDLINE records) and 700,000 to MEDLINE[2]. However, the methodological details of trials are often inadequately described by authors in the titles or abstracts, and not all records contain an abstract [4]. Due to these limitations, a wider (that is, more sensitive) search strategy is necessary to ensure articles are not missed, which leads to an imprecise dataset retrieved from electronic bibliographic databases. Typically, of the thousands of citations retrieved for a systematic review search over 90% are excluded on the basis of title and abstract screening [5].

Searching multiple databases is essential because different databases contain different records, and therefore, the coverage is widened. Also, searching multiple databases utilises differences in indexing to increase the likelihood of retrieving relevant items that are listed in several databases [6], but inevitably, this practice also retrieves overlapping content [7]. The degree of journal overlap estimated by Smith [8] over a decade ago indicated that about 35% of journals were listed in both MEDLINE and EMBASE. Journal overlap can vary from 10% to 75% [8,9,10,11,12] depending on medical speciality. More recently, the overlap in MEDLINE and EMBASE was found to be 79% [13] based on trials that had been included in 66 Cochrane systematic reviews.

The problem of overlapping content and subsequent retrieval of duplicate records is partially managed with commercial reference management software programs such as EndNote™[14], Reference Manager[15], Mendeley[16] and RefWorks[17]. They contain algorithms designed to identify and remove duplicate records using an auto-deduplication function. However, the detection of duplicate records can be thwarted by inconsistent citation details, missing information, or errors in the records. Typically, auto-deduplication is only partially successful [18], and the onerous task of manually sifting and removing the remaining duplicates rests with reviewers or information specialists. Removing such duplicates is an essential task to ensure systematic reviewers do not waste time screening the same citation multiple times. This study aimed to iteratively develop and test the performance of a new deduplication program against EndNote™ X6.

Methods

Systematic Review Assistant-Deduplication Module process of development

The Systematic Review Assistant-Deduplication Module (SRA-DM) project was developed in 2013 at the Bond University Centre for Research in Evidence-Based Practice (CREBP). The project aimed to reduce the amount of time taken to produce systematic reviews by maximising the efficiency of the various review stages such as optimising search strategies and screening, finding full text articles and removing duplicate citations.

The deduplication algorithm was developed using a heuristic-based approach with the aim of increasing the retrieval of duplicate records and minimising unique records being erroneously designated as duplicates. The algorithm was developed iteratively with each version tested against a benchmark dataset of 1,988 citations. Modifications were made to the algorithm to overcome errors in duplicate detection (Table 1). For example, errors often occurred due to variations in author names (e.g. first-name/surname sequence, use/absence of initialisation, missing author names and typographical errors), page numbers (e.g. full/truncated, or missing), text accent marks (e.g. French/German/Spanish) and journal names (e.g. abbreviated/complete, and '*the*' used intermittently). The performance of the SRA-DM algorithm was

compared with EndNote's default one step auto-deduplication process. To determine the reliability of SRA-DM, we conducted a series of validation tests with results of different literature searches (cytology screening tests, stroke and haematology) which were retrieved from searching multiple biomedical databases (Table 2).

Table 1. SRA-DM algorithm changes

Iterations	Changes to algorithms
First iteration	Matching criteria were based on simple field comparison (ignoring punctuation) with checks against the year field since this field has a lower probability for errors because it is restricted to integers 0–9 and therefore the best non-mistakable field.
Second iteration	Short format page numbers were converted to full format (e.g. 221–226, 221–6), and the algorithm was further modified to increase the sensitivity by incorporating matching criteria on authors OR title.
Third iteration	Match author AND title with the extension of the non-reference fields from only 'year' to year OR volume OR edition.
Fourth iteration	The fourth algorithm extended the matching criteria of the third algorithm, with the addition of an improved name matching system. This was context aware of author name variations, i.e. initialisation, punctuation and rearranged author listings using fuzzy logic, so that differences could be accommodated. For example, the following names are all syntactically equivalent and will match as identical authors:
	1. William Shakespeare
	2. W. Shakespeare
	3. W Shakespeare
	4. William John Shakespeare
	5. William J. Shakespeare
	6. W. J. Shakespeare
	7. W J Shakespeare
	8. Shakespeare, William
	9. Shakespeare, W
	10. Shakespeare, W, A
	11. Shakespeare, W, A, B, C
	12. William Shakespeare 1st
	13. William Shakespeare 2nd
	14. William Shakespeare IV
	15. William Adam Bob Charles Shakespeare XVI

Table 2 Databases searched for retrieval of citations for validation testing

Datasets	Databases searched
Cytology screening tests	1. Cochrane Controlled Trials Register (CCTR)
	2. Cochrane Database of Systematic Reviews (CDSR)
	3. EMBASE
	4. MEDLINE
	5. National Research Register (NRR)
	6. Database of Assessments of Reviews of Effectiveness
	7. NHS Health Technology Assessment (HTA)
	8. PreMEDLINE
	9. Science Citation Index
	10. Social Sciences Citation Index
Haematology dataset	1. MEDLINE
	2. EMBASE
	3. MEDLINE In-Process
	4. Biological Abstracts
	5. NHS Health Technology Assessment (HTA)
	6. Cochrane Controlled Trials Register (CCTR)
	7. Cochrane Database of Systematic Reviews (CDSR)
	8. CINAHL
	9. Science Citation Index
	10. Social Sciences Citation Index
Stroke dataset	1. MEDLINE
	2. EMBASE
	3. CENTRAL
	4. CINAHL
	5. PsycInfo

Definitions

A duplicate record was defined as being the same bibliographic record (irrespective of how the citation details were reported, e.g. variations in page numbers, author details, accents used or abridged titles). Where further reports from a single study

were published, these were not classed as duplicates as they are multiple reports which can appear across or within journals. Similarly, where the same study was reported in both journal and conference proceedings, these were treated as separate bibliographic records.

Testing against benchmark*

A total of 1,988 citations, derived from a search conducted on 29 July 2013 for surgical and non-surgical management for pleural empyema were used to test SRA-DM and EndNote™ X6. Six databases were searched (MEDLINE-Ovid, EMBASE-Elsevier, CENTRAL-Cochrane library, CINAHL-Ebasco, LILACS-Bireme, PubMed-NLM). To create the benchmark, citations were imported into EndNote™ database, sorted by author, inspected for duplicate records and manually coded as a unique or duplicate record; the database was reordered by article title and reinspected for further duplicates. Once the benchmark was finalised, duplicates were sought in EndNote™ using the default one-step auto-deduplication process which used the matching criteria of 'author', 'year' and 'title' (with the 'ignore spacing and punctuation' box ticked). A few additional duplicates were identified in EndNote™ and SRA-DM whilst cross-checking against the benchmark decisions, and the benchmark and results were updated to take account of these.

Data analysis

The accuracy of the results were coded against the benchmark according to whether it was a true positive (true duplicate, i.e. correctly identified duplicate), false positive (false duplicate, i.e. incorrectly identified as duplicate), true negative (unique record) or false negative (true duplicate, i.e. incorrectly identified as unique record). This process was repeated for results received after using the SRA-DM. Sensitivity is defined as the ability to correctly classify a record as duplicate and is the proportion of true positive records over the total number of records identified as true positive and false negative. Specificity is defined as the ability to correctly classify a record as being unique or non-duplicate and is the proportion of true negative records over the total number of records identified as true negative and false positive.

*For additional methodological details see Supplementary appendix B

Results

Training and development of SRA-DM

First and second iteration

The first iteration of the deduplication algorithm achieved 75.0% sensitivity and 99.9% specificity (Table 3). The matching criteria were based on field comparison (ignoring punctuation) with checks made against the year field. This field was chosen because the year field has a lower probability for errors since it is restricted to integers 0–9 and therefore is the best non-mistakable field. Eighty-four percent of undetected duplicates arose due to variations in page numbers (e.g. 221–226, 221–6). To address this, short format page numbers were converted to full format and the algorithm was further modified to increase the sensitivity by incorporating matching criteria on authors OR title. This increased the sensitivity of the second iteration to 95.7% with more duplicates detected, but as a consequence the number of false positives also increased (specificity 99.8%).

Table 3

Sensitivity† and specificity‡ of SRA-DM prototype algorithms and EndNote auto-deduplication (in a dataset of 1,988 citations, including 799 duplicates)

	Respiratory study				
	First iteration SRA-DM	**Second iteration SRA-DM**	**Third iteration SRA-DM**	**Fourth iteration SRA-DM**	**EndNote**
True positive (*n*) (correctly identified duplicates)	600	765	543	674	410
False negative (*n*) (duplicates missed)	199	34	256	125	391
Sensitivity (%)	*75.1*	*95.7*	*68.0*	*84.4*	*51.2*
True negative (*n*) (correctly identified unique records)	1,188	1,186	1,189	1,189	1,185
False positive (*n*) (incorrectly identified as duplicates)	1	3	0	0	2
Specificity (%)	*99.9*	*99.8*	*100.0*	*100.0*	*99.8*

$$\dagger \text{Sensitivity} = \frac{\text{number of true positive results}}{\text{number of true positives + number of false negatives}};$$

$$\ddagger \text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives + number of false positives}}$$

Third iteration

The third iteration was modified to match author AND title with the extension of the non-reference fields from only 'year' to year OR volume OR edition. This distinguished the references that were similar (e.g. same author and title combination) but contained different source publications, and this improved the specificity to 100% but the sensitivity was reduced (68.0%).

Fourth iteration

The fourth iteration was modified to accommodate author name variations using fuzzy logic so that differences in names spelt in full or initialised, differences in the

ordering of name and different punctuation could be accommodated (Table 1); this increased the sensitivity to 84.4% by correctly identifying 674 citations as duplicates (TP), 1,189 citations as unique records (TN), no false positives occurred (100% specificity) and only 125 duplicate records were undetected (FN). This fourth iteration of SRA-DM was then compared against EndNote™. EndNote™ identified 412 of the 1,988 citations as duplicates. Of these, 410 were correctly identified as duplicates (TP) and two were incorrectly designated as duplicates (FP), and 1,185 citations were correctly identified as unique records (TN) and 391 duplicate citations were undetected (FN). The sensitivity of EndNote™ was 51.2% and specificity 99.8%. Compared with EndNote™, SRA-DM produced a 64% increase in sensitivity and no loss of specificity.

Validation results

The fourth iteration of SRA-DM was further tested with three additional datasets using search topics from cytology screening tests (*n* = 1,856), stroke (*n* = 1,292) and haematology (*n* = 1,415) (Table 2). These were obtained from existing searches performed by information specialists to widen the scope of the validation tests. SRA-DM algorithm was consistently more sensitive (Table 4) at detecting duplicates than EndNote™ [cytology screening: 90% vs 63%; stroke: 84% vs 73% and haematology: 84% vs 64%] and specificity of SRA-DM was 100% accurate, i.e. no false positives occurred. In contrast, the average specificity of EndNote™ was lower (99.7). These false positives occurred in EndNote™ due to citations with the same authors and title being published in other journals or as conference proceeding. Compared with EndNote™, the average percentage increase in duplicates detected by SRA-DM across all four bibliographic searches was 42.8%.

Table 4

Sensitivity† and specificity‡ of SRA-DM and EndNote auto-deduplication (validation testing)

	Cytology screening		Stroke		Haematology	
	SRA-DM	**EndNote**	**SRA-DM**	**EndNote**	**SRA-DM**	**EndNote**
True positive (correctly identified duplicates)	1,265	885	426	372	208	159
False negative (duplicates missed)	139	518	81	134	38	87
Sensitivity (%)	*90.10*	*63.08*	*84.02*	*73.52*	*84.55*	*64.63*
True negative (correctly identified unique records)	452	452	785	784	1,169	1,165
False positive (incorrectly identified duplicates)	0	1	0	2	0	4
Specificity (%)	*100.00*	*99.78*	*100.00*	*99.75*	*100.00*	*99.66*

$$\text{Sensitivity} = \frac{\text{number of true positive results}}{\text{number of true positives} + \text{number of false negatives}}$$

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

Discussion

Our findings demonstrated that SRA-DM identifies substantially more duplicate citations than EndNote™ and has greater sensitivity [(84% vs 51%), (90% vs 63%), (84% vs 73%), (84 vs 64%)]. The specificity of SRA-DM was 100% with no false positives, whereas the specificity of EndNote™ was imperfect.

Waste in research occurs for several methodological, legislative and reporting reasons [19, 20, 21, 22]. Another form of waste is inefficient labouring, in part, because of non-standardised citations details across bibliographic databases, perfunctory error checking and absence of a unique trial identification number for it and its associated further multiple reports. If these issues were solved at source, manual duplicate checking would be unnecessary. Until these issues are resolved, deploying the SRA-DM will save information specialists and reviewers valuable time by identifying on average a further 43% of duplicate records.

Several citations were wrongly designated as duplicates by EndNote™ auto-deduplication due to different citations sharing the same authors and title but published in other journals or as conference proceedings. In a recent study by Jiang [23], the authors also found that EndNote™, for the same reason, had erroneously assigned unique records as duplicates. It is probable that in most scenarios no important loss of data would occur; although sometimes additional methodological or outcome data are reported, and ideally these need to be retained for inspection. A recent study by Qi [18] examined the content of undetected duplicate records in EndNote™ and found that errors often occurred due to missing or wrong data in the fields, especially for records retrieved from EMBASE database. This also affected the sensitivity of SRA-DM, with duplicates undetected due to missing or wrong or extraneous data in the fields.

During the training and development stage, the four iterations of SRA-DM achieved sensitivities ranging from 68%, 75%, 84% and 96% with the most sensitive (96%) achieved with a trade-off in specificity (99.75%) with three false positives. For systematic reviews and Health Technology Assessment reports, the aim is to

conduct comprehensive searches to ensure all relevant trials are identified [24]; thus, losing even three citations is undesirable. Therefore, the final algorithm (fourth iteration) with the lower sensitivity (84%) but perfect (100%) specificity was preferred. Future developments with SRA-DM may incorporate two algorithms, first using the 100% specific algorithm to automatically remove duplicates and another algorithm with higher sensitivity (albeit with lower specificity) to identify the remaining duplicates for manual verification. If this strategy was implemented on the respiratory dataset using the fourth and second algorithm (Table 3), only 91 out of 1,988 citations would have to be manually checked and only 34 duplicates would remain undetected.

In spite of this major improvement with the SRA-DM, no software can currently detect all duplicate records, and the perfect uncluttered dataset remains elusive. Undetected duplicates in SRA-DM occurred due to discrepancies such as missing page numbers or too much variance with author names. Duplicates were also missed because the OVID MEDLINE platform inserted additional extraneous information into the title field (e.g. [Review] [72 refs]) whereas the same article retrieved from EMBASE or other non-OVID MEDLINE platforms (i.e. PubMed, Web of Knowledge) report only the title. Some of these problems could be overcome in the future with record linkage and citation enrichment techniques to populate blank fields with meta-data to increase the detection rate.

Strengths and weaknesses

The deduplication program was developed to identify duplicate citations from biomedical databases and has not been tested on other bibliographic records such as books and governmental reports and therefore may not perform as well with other bibliographies. However, the deduplication program was developed iteratively to remove problems of false positives and was tested on four different datasets which included comprehensive searches using 14 different databases that are used by information specialists, and therefore, similar efficiencies should occur in other medical specialities. Also, the accuracy of SRA-DM was consistently higher than that of EndNote™, and these finding are probably generalizable to other biomedical database searches due to the same records types and fields used. It is possible that

some duplicates were not detected during the manual benchmarking process, although the database was screened twice first by author and then by title, and additional cross-checking was performed by manually comparing the benchmark against EndNote™ auto-deduplication and SRA-DM decisions—thus minimising the possibility of undetected duplicates.

Whilst we compared SRA-DM against the typical default EndNote™ deduplication setting, we recognise that some information specialists adopt additional steps whilst performing deduplication in EndNote™. For example, they may employ multi-stage screening or attempt to replace incomplete citations by updating citation fields with the 'Find References Update' feature in EndNote™. However, many researchers and information specialists do not employ such techniques, and our aim was to address deduplication with an automated algorithm and compare it against the default deduplication process in EndNote™. Qi [18] recommended employing a two-step strategy to address the problem of undetected duplicates by first performing auto-deduplication in EndNote™ followed by manual hand screening to identify remaining duplicates. This basic strategy is used by some information specialists and systematic reviewers but is inefficient due to the large proportion of unidentified duplicates. Other more complex multi-stage screening strategies have been suggested [25] but are EndNote-specific and not viable for other reference management software.

Conclusions

The deduplication algorithm has greater sensitivity and specificity than EndNote™. Reviewers and information specialists incorporating SRA-DM into their research procedures will save valuable time and reduce resource waste. The algorithm is open source [26] and the SRA-DM program is freely available to users online[27]. It allows similar file manipulation to EndNote™ and currently accepts XML, RIS and CSV file formats enabling citations to be exported directly to RevMan software. It has the option of automatic duplicate removal or manual pair-wise duplicate screening performed individually or with a co-reviewer.

Declarations

Sources of funding

NHMRC Australia Fellowship: GNT0527500.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed to the study concept and design. JR devised the testing and analysis of the algorithms. MC wrote and revised the algorithm codes. JR drafted the initial manuscript. TH, PG and MC contributed to the manuscript and all the revisions. All authors read and approved the final manuscript.

References

1. Islamaj Dogan R, Murray GC, Névéol A, Lu Z: **Understanding PubMed user search behavior through log analysis.** Database J Biol Databases Curation 2009, **2009:**1.
2. **MEDLINE - fact sheet**. [http://www.nlm.nih.gov/pubs/factsheets/medline.html]
3. **Embase**. [http://www.ovid.com/webapp/wcs/stores/servlet/ProductDisplay?storeId=13051&catalogId=13151&langId=-1&partNumber=Prod-903]
4. Lefebvre C, Eisinga A, McDonald S, Paul N: **Enhancing access to reports of randomized trials published world-wide–the contribution of EMBASE records to the Cochrane central register of controlled trials (CENTRAL) in the Cochrane library.** Emerg Themes Epidemiol 2008, **5:**13. 10.1186/1742-7622-5-13
5. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH: **Semi-automated screening of biomedical citations for systematic reviews.** BMC Bioinformatics 2010, **11:**1–11. 10.1186/1471-2105-11-1
6. Sampson M, McGowan J, Cogo E, Horsley T: **Managing database overlap in systematic reviews using batch citation matcher: case studies using scopus.** J Med Libr Assoc 2006, **94:**461–463.
7. Sievert MC, Andrews MJ: **Indexing consistency in information science abstracts.** J Am Soc Inf Sci 1991, **42:**1–6. 10.1002/(SICI)1097-4571(199101)42:1<1::AID-ASI1>3.0.CO;2-9
8. Smith B, Darzins P, Quinn M, Heller R: **Modern methods of searching the medical literature.** Med J Aust 1992, **2:**603–611.
9. Kleijnen J, Knipschild P: **The comprehensiveness of MEDLINE and Embase computer searches. Searches for controlled trials of homoeopathy, ascorbic acid for common cold and ginkgo biloba for cerebral insufficiency and intermittent claudication.** Pharm Weekbl Sci 1992, **14:**316–320. 10.1007/BF01977620
10. Odaka T, Nakayama A, Akazawa K, Sakamoto M, Kinukawa N, Kamakura T, Nishioka Y, Itasaka H, Watanabe Y, Nose Y: **The effect of a multiple**

literature database search–a numerical evaluation in the domain of Japanese life science. J Med Syst 1992, **16:**177–181. 10.1007/BF00999380

11. Rovers JP, Janosik JE, Souney PF: **Crossover comparison of drug information online database vendors: dialog and MEDLARS.** Ann Pharmacother 1993, **27:**634–639.
12. Ramos-Remus C, Suarez-Almazor M, Dorgan M, Gomez-Vargas A, Russell AS: **Performance of online biomedical databases in rheumatology.** J Rheumatol 1994, **21:**1912–1921.
13. Royle P, Milne R: **Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches.** Int J Technol Assess Health Care 2003, **19:**591–603.
14. **EndNote**. [http://endnote.com/]
15. **Reference manager**. [http://www.refman.com/]
16. **Mendeley**. [http://www.mendeley.com/]
17. **RefWorks**. [www.refworks.com]
18. Qi X, Yang M, Ren W, Jia J, Wang J, Han G, Fan D: **Find duplicates among the PubMed, EMBASE, and Cochrane library databases in systematic review.** PLoS One 2013, **8:**e71838. 10.1371/journal.pone.0071838
19. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E: **Reducing waste from incomplete or unusable reports of biomedical research.** Lancet 2014, **383:**267–276. 10.1016/S0140-6736(13)62228-X
20. Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, Krumholz HM, Ghersi D, van der Worp HB: **Increasing value and reducing waste: addressing inaccessible research.** Lancet 2014, **383:**257–266. 10.1016/S0140-6736(13)62296-5
21. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, Howells DW, Ioannidis JP, Oliver S: **How to increase value and reduce waste when research priorities are set.** Lancet 2014, **383:**156–165. 10.1016/S0140-6736(13)62229-1

22. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R: **Increasing value and reducing waste in research design, conduct, and analysis.** Lancet 2014, **383:**166–175. 10.1016/S0140-6736(13)62227-8

23. Jiang Y, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR: **Rule-based deduplication of article records from bibliographic databases.** Database (Oxford) 2014, **2014:**1–7.

24. **Cochrane handbook for systematic reviews of interventions**. [http://www.cochrane.org/handbook]

25. **Removing duplicates in retrieval sets from electronic databases: comparing the efficiency and accuracy of the Bramer-method with other methods and software packages**. [http://www.iss.it/binary/eahi/cont/57_Wichor_M._Bramer.pdf]

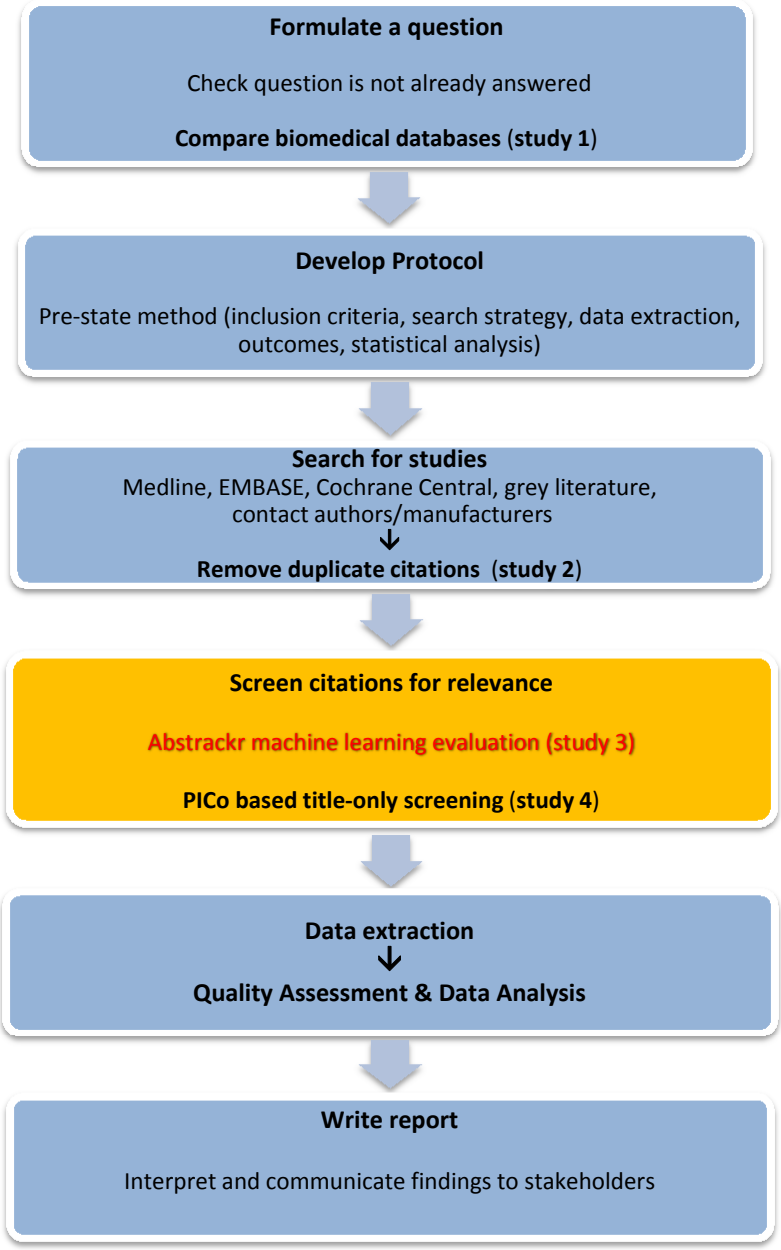
26. **Source code**. [https://github.com/CREBP/SRA]

27. **Systematic review assistant - deduplication module**. [http://crebp-sra.com]

Copyright

Chapter 5

Semi-automated citation screening



Key steps for conducting a systematic review and where studies for this PhD are focussed

The previous chapter highlighted the problem of, and potential partial solution to, duplicate records retrieved from systematic searches of databases. The next stage of systematic reviewing, once the duplicate records are removed, requires screening the titles and abstracts of records to identify relevant studies for inclusion. Title and abstract screening is time-consuming for researchers and previous attempts at applying text mining to screening records have been inadequate because a threshold of 95% retrieval was used as an acceptable measure of success to identify relevant records. For systematic review purposes, this threshold is too low and would be unacceptable for commissioning bodies since 5% loss of data would potentially bias the findings. More recently, semi-automated screening methods have been developed specifically for systematic reviews evaluated against higher thresholds of accuracy, but their suitability for systematic reviewing remained unclear due to limited research and lack of independent evaluation.

In this chapter, the role of semi-automated screening is introduced and the current state of the technology. The limited evaluation surrounding text mining for systematic reviews and the absence of independent evaluation of existing text mining tools led to developing a research study evaluating the predictive screening software Abstrackr. The merits and demerits of text mining with Abstrackr are discussed and compared with the screening accuracy of manual screening and the diminishing returns of text mining with different systematic review topics.

Summary

The published study demonstrated that semi-automated screening with Abstrackr has the potential to reliably identify relevant citations through predictive screening and reduce workload from 9 to 80%. Nonetheless, in two datasets a small proportion of relevant abstracts were incorrectly predicted as irrelevant by Abstrackr and therefore caution is needed using Abstrackr as a stand-alone application. After the research was published in the journal, 'Systematic Reviews' the article received much interest from academic researchers and was listed as one of the most influential systematic review articles read in 2015[43].

5.1 Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers

Systematic Reviews (2015) **4**:80. DOI: 10.1186/s13643-015-0067-6

John Rathbone

Tammy Hoffmann

Paul Glasziou

Abstract

Background

Citation screening is time consuming and inefficient. We sought to evaluate the performance of Abstrackr, a semi-automated online tool for predictive title and abstract screening.

Methods

Four systematic reviews (aHUS, dietary fibre, ECHO, rituximab) were used to evaluate Abstrackr. Citations from electronic searches of biomedical databases were imported into Abstrackr, and titles and abstracts were screened and included or excluded according to the entry criteria. This process was continued until Abstrackr predicted and classified the remaining unscreened citations as relevant or irrelevant. These classification predictions were checked for accuracy against the original review decisions. Sensitivity analyses were performed to assess the effects of including case reports in the aHUS dataset whilst screening and the effects of using larger imbalanced datasets with the ECHO dataset. The performance of Abstrackr was calculated according to the number of relevant studies missed, the workload saving, the false negative rate, and the precision of the algorithm to correctly predict relevant studies for inclusion, i.e. further full text inspection.

Results

Of the unscreened citations, Abstrackr's prediction algorithm correctly identified all relevant citations for the rituximab and dietary fibre reviews. However, one relevant citation in both the aHUS and ECHO reviews was incorrectly predicted as not relevant. The workload saving achieved with Abstrackr varied depending on the complexity and size of the reviews (9% rituximab, 40% dietary fibre, 67% aHUS, and 57% ECHO). The proportion of citations predicted as relevant, and therefore, warranting further full text inspection (i.e. the precision of the prediction) ranged from 16% (aHUS) to 45% (rituximab) and was affected by the complexity of the reviews. The false negative rate ranged from 2.4 to 21.7%. Sensitivity analysis performed on the aHUS dataset increased the precision from 16 to 25% and increased the workload saving by 10% but increased the number of relevant studies

missed. Sensitivity analysis performed with the larger ECHO dataset increased the workload saving (80%) but reduced the precision (6.8%) and increased the number of missed citations.

Conclusions

Semi-automated title and abstract screening with Abstrackr has the potential to save time and reduce research waste.

Background

Systematic reviews require a comprehensive search and appraisal of the literature to identify all relevant studies for inclusion. Typically, this involves a team of reviewers inspecting thousands of records that are produced from database searches. The large number of citations retrieved is partly due to the inadequate coding of studies indexed in biomedical databases such as MEDLINE and EMBASE. This produces imprecise search results; sometimes less than 1% of studies screened are included in a systematic review [1, 2]. Systematic reviews have also become more time consuming due to the growth in the volume and scatter of randomised trials [3], additional reporting steps [4, 5, 6], and the incorporation of more complex methodologies such as network meta-analysis and the acquisition of clinical study reports [7]. Consequently, many systematic reviews are out of date [8, 9]. With all these challenges, there is a need to adopt techniques from computer science that can semi-automate screening in order to expedite the process of study selection.

Text mining techniques are used to identify relevant information from text using statistical pattern learning that recognises patterns in data. Typically, such patterns are learnt from labelled training data that are then applied to datasets. A common application of such techniques is used to separate spam from real emails. Pattern recognition algorithms aim to provide the most likely matching of the inputs, taking into account their statistical variation. They have been applied in a variety of ways in evidence-based medicine to expedite tasks that would otherwise be omitted due to the time and cost involved if they were performed manually. For example, text mining has been used to assess the frequency of adverse effects of drugs by

analysing patient medical records [10] and to expedite scoping searches [11]. Text mining has the potential to reduce the workload of systematic reviewers by assisting with the identification of relevant trials during the title and abstract screening stage of a systematic review.

Abstrackr [12] is a free, open-source [13], citation screening program, currently at beta testing stage that uses an algorithm within an active learning framework to predict the likelihood of citations being relevant. It uses text unigrams and bigrams within the annotated abstracts for the predictive modelling. Abstrackr biases the citations so that the most relevant are prioritised for screening first. Only limited research to date has been conducted into the strengths and limitations of semi-automated citation screening [14, 15]. The aim of this study was to evaluate the performance of the Abstrackr algorithm. It was chosen for evaluation in preference to other text mining tools because existing literature indicates that the recall accuracy of Abstrackr is very high [14, 15, 16, 17], and therefore, a promising predictive text mining tool for systematic reviews, where the primary goal is to identify all relevant studies.

Methods

Four systematic review datasets derived from the literature searches of completed systematic reviews [1, 18, 19, 20] were used to evaluate Abstrackr. Three systematic reviews evaluated treatment effectiveness: dietary fibre interventions for colorectal cancer, rituximab and adjunctive chemotherapy interventions for non-Hodgkin's lymphoma, eculizumab for atypical hemolytic uremic syndrome (aHUS), and one diagnostic accuracy review of echocardiography (ECHO) was included. Each systematic review was chosen because of their different characteristics: for example, the aHUS review included all study designs except case reports; the interventions in the rituximab review included multiple chemotherapy interventions rather than a simple drug A versus drug B comparison; the dataset from the dietary fibre review was from a specialised register which provides a more homogeneous and smaller set of citations and therefore presents a challenge to supervised machine learning algorithms because they perform better on large datasets; and the ECHO was chosen because it is a diagnostic accuracy review*.

Citations were uploaded to Abstrackr, and titles and abstracts were screened for relevance by one author with relevant studies selected for inclusion and clearly irrelevant studies excluded. Screening continued until the algorithm's stopping criterion indicated that predictions were available for viewing. This is based upon a simple heuristic requiring a set number of citations to be screened manually. The remaining unscreened citations were inspected according to the probability estimates and hard binary prediction made by the algorithm and cross-checked against the original review decisions.

The performance of Abstrackr was assessed by calculating the precision, the false negative rate, the proportion missed, and the workload saving. The precision is the percentage of citations predicted relevant by Abstrackr that are subsequently deemed relevant by the *reviewer* for further full text inspection. The false negative rate is the percentage of citations that are relevant for further full text inspection but were predicted to be irrelevant by Abstrackr. The proportion missed is the number of studies missed by Abstrackr that were included in the published reviews, out of those studies predicted to be irrelevant. The workload saving is the proportion of citations predicted irrelevant out of the total number of citations.

A post hoc sensitivity analysis was performed on the aHUS dataset because many of the included and excluded studies were methodologically similar, and therefore, excluding near matches might impede the learning algorithm. For example, case reports were originally excluded, but case series and RCTs were relevant and included. Therefore, by rescreening the aHUS dataset and also including case reports, we sought to determine if their inclusion would improve the machine learning precision by reducing superficially conflicting decisions. A post hoc sensitivity analysis was also performed on a substantially larger ECHO dataset to determine if this would affect the workload saving.

---

*For supplementary methodological details see appendix C

Results

A total for four datasets from existing systematic reviews (aHUS *n* = 1415), (dietary fibre *n* = 517), (ECHO *n* = 1735) and (Rituximab *n* = 1042) were uploaded to Abstrackr and screened for relevance until the classification algorithm made predictions.

Atypical haemolytic uremic syndrome dataset (excluding case reports)

Of 1415 citations, 251 citations were screened (18%) before Abstrackr made the predictions, leaving 1164 (82%) citations unscreened. Of these, Abstrackr predicted that 374 citations were potentially relevant, and 63 were found to be relevant, giving a precision of 16.8% (Fig. 1). The false negative rate was 10% (Fig. 2). Of the 790 citations predicted not relevant, one citation was included in the review, giving a percentage missed of 0.13% (Fig. 3). As 44% of citations required screening and checking for relevance, a workload saving of 56% was achieved (Fig. 4).

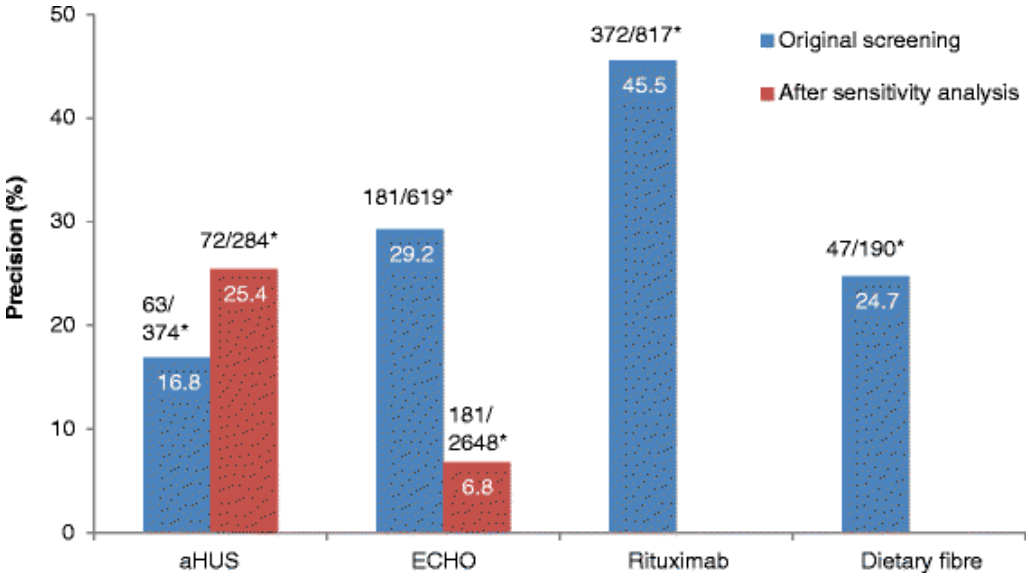


Figure 1. Percentage of citations predicted by Abstrackr that were relevant for further full text inspection. *Raw numbers of the proportion of citations selected for inspection

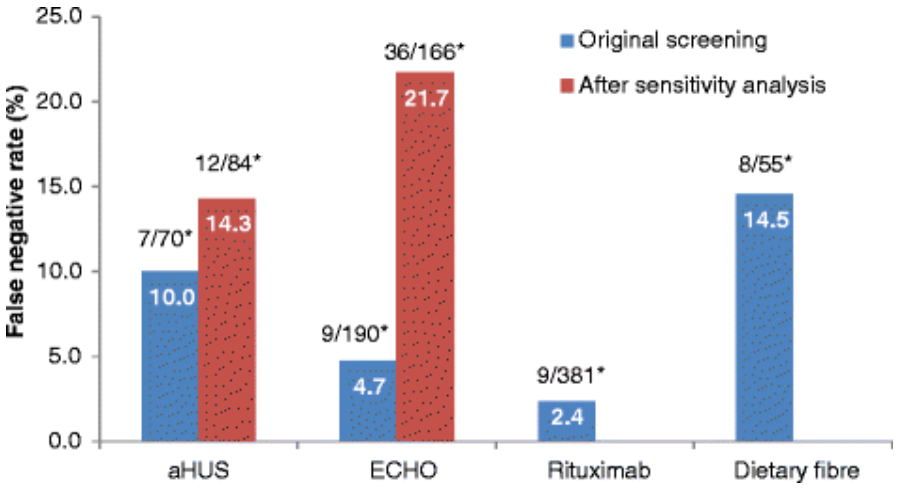


Figure 2. False negative rate. *Raw numbers of the proportion of citations incorrectly predicted by Abstrackr to be irrelevant for further inspection

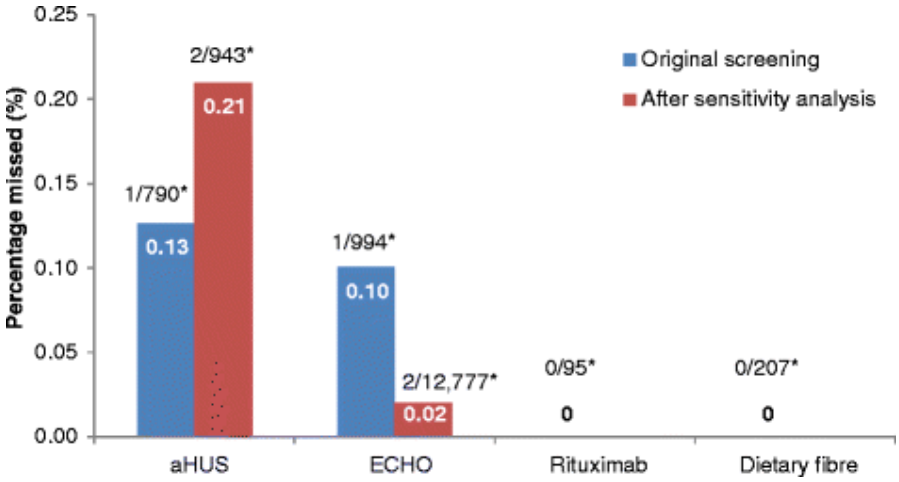


Figure 3. Percentage of studies missed by Abstrackr—but were included in the reviews. *Raw numbers of the proportion of citations missed (predicted not relevant)

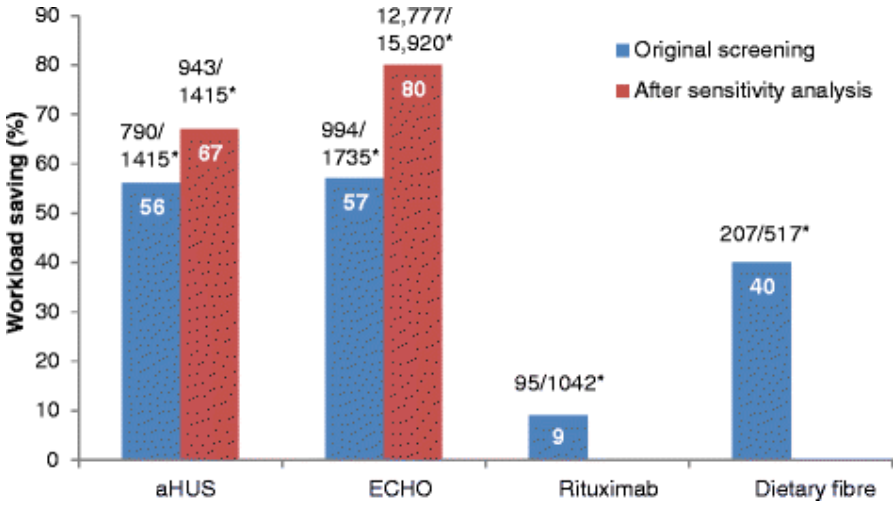


Figure 4. Workload saving (%) when using Abstrackr in each of the four datasets. *Raw numbers of the proportion of citations predicted not relevant from the total

Sensitivity analysis of atypical haemolytic uremic syndrome dataset (including case reports)

The citations were re-screened using the same decisions to include or exclude citations—with the exception that case reports were included (even though irrelevant). This 'homogeneous' screening method increased the precision from 16.8 to 25.4% (Fig. 1) and the false negative rate was 14.3% (Fig. 2). The number of relevant citations missed, however, increased to two citations (0.21%) (Fig. 3). The workload saving increased from 56 to 67% (Fig. 3).

Dietary fibre for colorectal cancer dataset

Of 517 citations, 120 citations (23%) were screened before Abstrackr made predictions. Abstrackr predicted a further 190 were potentially relevant, and 47 were found to be relevant, giving a precision of 24.7% (Fig. 1). The false negative rate was 14.5% (Fig. 2). Of the remaining 207 citations predicted as not relevant by Abstrackr, none were included in the review—giving a 0% missed (Fig. 3). Sixty percent of citations required screening and checking for relevance, providing a workload saving of 40% (Fig. 4).

Echocardiography for stroke dataset

Of 1735 citations, 122 (7%) were screened before Abstrackr made predictions. Abstrackr predicted that a further 619 were potentially relevant, and 181 were found to be relevant giving a precision of 29.2% (Fig. 1). The false negative rate was 4.7% (Fig. 2). Of the remaining 994 citations predicted as not relevant by Abstrackr, 993 were correctly excluded; however, one citation that was included in the review was missed, giving a percentage missed of 0.10% (Fig. 3). Forty-three percent of citations required screening and checking for relevance, providing a workload saving of 57% (Fig. 4).

Sensitivity analysis of echocardiography for stroke (large dataset)

The citations were re-screened using a larger dataset of 15,920 citations to determine if precision and workload saving were affected. Abstrackr made predictions after 495 citations were screened and predicted that 2648 citations were potentially relevant. Of these, 181 were found to be relevant for full text inspection, giving a precision of 6.8%. The false negative rate was 21.7% (Fig. 2). Of the remaining 12,777 predicted as not relevant by Abstrackr, 12,775 were correctly predicted as not relevant. However, two citations that were included in the published review were missed, giving a percentage missed of 0.02%. Twenty percent of citations required screening, providing a workload saving of 80% (Fig. 4).

Rituximab for Non-Hodgkins lymphoma

Of 1042 citations, 130 citations (12%) were screened before Abstrackr made predictions. Abstrackr predicted 817 citations were potentially relevant, and 372 were found to be relevant giving a precision of 45.5% (Fig. 1). The false negative rate was 2.4% (Fig. 2). Of the remaining 95 citations predicted as not relevant by Abstrackr, none were included in the review, giving a percentage missed of 0 (Fig. 3). As 91% of citations required screening and checking for relevance, there was only a 9% workload saving (Fig. 4).

Discussion

This study found that Abstrackr has the potential to reliably identify relevant citations and reduce workload from 9 to 80%. In two datasets, all relevant citations were identified, and in the other two datasets, only one citation was missed. The false negative rate ranged from 2 to 21%. Overall, precision was good although affected by the complexity of the review.

In the aHUS dataset, precision was only 16.8%. This was due to the complexities of the study inclusion criteria which included case series as well as other higher quality study designs but not case reports that were excluded during screening. Because of the lexical similarity between case reports and case series, excluding case reports introduced greater variance into the machine learning algorithm with apparent conflicting decisions and consequently reduced precision. The sensitivity analysis demonstrated that by reducing 'noise', the precision could be increased. This problem of 'noise' with machine learning is common [21], and one strategy to increase precision during the data-training phase is to include close matching records [2], to ensure the active learning algorithm is not adversely affected, although this requires a degree of expertise to make decisions contrary to the PICOS (Participants, Interventions, Comparators, Outcomes and Study design) inclusion criteria. The ECHO sensitivity analysis had the worst precision (6.8%) because of the 15,920 citations wherein there was only about 0.9% that was relevant. Such imbalanced datasets are problematic for supervised machine learning models like Abstrackr, because the predictions are biased towards the majority non-relevant class at the expense of the minority-relevant class [22] and therefore produce many falsely weighted predictions, i.e. irrelevant citations. Nevertheless, this was off-set by the considerable workload saving.

The false negative rates ranged from 2 to 21.7% and represent the percentage of citations that were relevant for further full text inspection but were predicted to be irrelevant by Abstrackr and were therefore 'missed'. However, the actual percentage missed were in the range of 0 to 0.21% and represent the true final proportion of citation missed by Abstrackr that were included in the review. Therefore, the classification model was almost completely reliable. The citation missed from the

aHUS and ECHO datasets did not contain an abstract, only a title and therefore the probability of being predicted relevant was reduced. The aHUS sensitivity analysis missed two citations, and both contained no abstract. The ECHO sensitivity analysis missed two citations, one without an abstract, whilst the other did contain an abstract and it is unclear why this citation was not detected as relevant. However, these problems could be minimised by retaining citations without an abstract for manual inspection.

The complexity of the review PICOS criteria also affected the workload saving. The workload saving in the rituximab dataset was low (9%) due to the rituximab intervention having multiple adjunctive chemotherapy treatments which overlapped with non-relevant studies. Therefore, the good precision and perfect recall accuracy with the rituximab data were off-set by the minimal workload savings suggesting that complex reviews may be less suited to semi-automated screening. Nevertheless, the average workload saving across the four datasets was 41% and is similar to the findings reported by the developers of Abstrackr who achieved a 40% saving in workload from two datasets[14].

Other data mining algorithms have achieved similar (40%) workload savings [16] but recall (identifying relevant records) was lower (90 to 95%), partly because testing was performed on datasets without a specifically associated research question. This makes comparisons with the results of this study difficult. Whilst another text mining algorithm [17] achieved workload savings ranging from 8.5 to 62% with 15 test datasets, which are similar to our findings with Abstrackr (9 to 80%), their results were based on a threshold of a minimum 95% recall of relevant studies, which is too low for systematic reviews. The developers of Abstrackr reported a recall accuracy of 100% for relevant studies from three genetics-related datasets and 99% for a fourth dataset, whilst the average specificity across the four datasets was 87% [14]. Their results were based on training the algorithm with balanced datasets, which have a similar number of relevant and irrelevant trials from the original systematic review, and using this trained algorithm to automatically find studies for the updates of the genetics-based systematic reviews. This approach is noteworthy since systematic reviews often require update searches to be performed within 2 years of the first published version [23], therefore, implementing this strategy, by retaining

the original classification model, would expedite the process of updating systematic reviews.

Strengths and weaknesses of the research

Our findings may be limited by the four datasets used, and citations from other clinical specialities may yield different precision and workload saving and miss more relevant studies for inclusion, especially if the title and abstract descriptions are inadequate or the study designs are more complex. Our datasets were from recently published systematic reviews that included trials published mostly from 1995 onwards, and therefore, may contain better descriptions than older trials that were published before the CONSORT [24] reporting guidelines were introduced in 1996. Nevertheless, our results for identifying relevant trials are similar to the high recall results of Wallace (2010 and 2012) and indicate that similar accuracy could be achieved when using other datasets of medical citations. Previous text mining studies have mainly evaluated performance in terms of recall and specificity; however, our results also analysed the precision of the predictive model since this measures how precisely the algorithm selects studies for further full text inspection and mirrors the working steps of a systematic reviewer. Precision, however, is subjective and influenced by the reviewer's expertise which can affect their screening judgements. The ECHO sensitivity analysis demonstrated that workload saving with semi-automated screening is more pronounced with large datasets, and therefore, greater savings could have resulted had we screened larger reviews. Nonetheless, the results provide a reasonable estimate of the algorithm's typical performance during semi-automated screening.

This study and others that have evaluated semi-automated screening with support vector models [14, 15], semantic vector models [16], and complement naïve Bayes models [17] indicate that considerable workload savings can be achieved. The ability to identify all relevant citations with Abstrackr was very high but imperfect. Such accuracy, however, is acceptable as a stand-alone tool for scoping searches and non-systematic reviews where not every published study needs to be included. It is noteworthy, however, that human citation screening is imperfect with relevant studies wrongly excluded [25]. Given that Abstrackr's inaccuracy is similar to a

human screener, it could be utilised as the second screener. Abstrackr's classification prediction model uses a somewhat arbitrary cut-off point at which the proportion of citations screened triggers the algorithm prediction. However, this suggests that an adjustable stopping heuristic could be used, so accuracy could be further improved albeit with the trade-off that more citations are screened during the training phase.

Future developments with semi-automated screening would benefit from retaining the original classification model developed during the original review, so future systematic review updates may be screened automatically without the re-input of a reviewer. Abstrackr is not currently a consumer level product, and only the unscreened citations (the predictions) are exportable with only the title bibliographic details made available, and further developments are needed to create a fully integrated application that systematic reviewers and information specialists can use.

Text mining algorithms have been proposed [26] to improve automated screening by including keywords to bias the predictive classification model so that citations containing such keywords are more likely to be identified. This approach could be further aided by citation enrichment. For example, keywords of high relevance such as the PICOS details should improve the recall accuracy of semi-automated screening algorithms (and trial searching). Enriching citations is already being used for the EMBASE project [27] by coding citations with the type of study design through crowd sourcing. Further research and innovations in this underexplored area is needed to advance current methods, and eventually enable semi-automated screening to fully replace manual screening. Current text mining research [28] is focused on advancing screening retroactively and is restrained by the limitations of the data available. A more successful approach may require collaboration with biomedical database providers to ensure that citations are adequately labelled prospectively and retrospectively using strategies such as record linkage techniques, crowd sourcing, or access to a central repository, whereby PICOS details can be inputted and linked to all bibliographic databases.

Conclusions

Semi-automated screening with Abstrackr can potentially expedite the title and abstract screening phase of a systematic review. Although the accuracy is very high, relying solely on its predictions when used as a stand-alone tool is not yet possible. Nevertheless, efficiencies could still be attained by using Abstrackr as the second reviewer thereby saving time and resources.

Abbreviations

aHUS: atypical hemolytic uremic syndrome

CENTRAL: The Cochrane Central Register of Controlled Trials

CINAHL: Cumulative Index to Nursing and Allied Health Literature

CONSORT: Consolidated Standards of Reporting Trials

ECHO: echocardiography

EMBASE: Excerpta Medica Database

MEDLINE: Medical Literature Analysis and Retrieval System Online

PICOS: Participants, Interventions, Comparators, Outcomes, Study design

Declarations

Acknowledgements

We would like to acknowledge the invaluable technical support provided by Byron Wallace during testing.

Sources of funding

NHMRC Australia Fellowship: GNT05275

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed to the study concept and design. JR devised the testing and analysis of the Abstrackr algorithm and drafted the initial manuscript. TH and PG contributed to the manuscript and all the revisions. All authors read and approved the final manuscript.

References

1. Rathbone J, Kaltenthaler E, Richards A, Tappenden P, Bessey A, Cantrell A. **A systematic review of eculizumab for atypical haemolytic uraemic syndrome (aHUS)**. BMJ Open. 2013;3:1–11.
2. Frunza O, Inkpen D, Matwin S. **Building systematic reviews using automatic text classification techniques.** Stroudsburg, PA, USA: Proceedings of the 23rd International Conference on Computational Linguistics; 2010.
3. Hoffmann T, Erueti C, Thorning S, Glasziou P. **The scatter of research: cross sectional comparison of randomised trials and systematic reviews across specialties**. BMJ. 2012;344:e3223.
4. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. 2009. **Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement**. BMJ. 2009;339:b2535. The PRISMA Statement. http://www.prisma-statement.org/statement.htm.
5. GRADEpro. 2015. http://www.guidelinedevelopment.org/. Accessed 2014.
6. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. **The Cochrane Collaboration's tool for assessing risk of bias in randomised trials**. BMJ. 2011;343:d5928.
7. Jefferson T, Jones MA, Doshi P, Del Mar CB, Hama R, Thompson MJ et al. **Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children**. Cochrane Database Syst Rev 2014, Issue 4.
8. Kristiansen I. **How up-to-date are Cochrane reviews?** Lancet. 2008; 371:384.
9. Shojania K, Sampson M, Ansari M, Ji J, Doucette S, Moher D. **How quickly do systematic reviews go out of date? A survival analysis**. Ann Intern Med. 2007;147:224–33.
10. Leeper N, Bauer-Mehren A, Iyer S, Lependu P, Olson C, Shah N. **Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes**. PLoS ONE. 2013;8:e63499.

11. Shemilt I, Antonia A, Hollands G, Marteau T, Ogilvie D, O'Mara-Eves A, et al. **Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews**. Res Synth Methods. 2013;5:31–49. Abstrackr. http://abstrackr.cebm.brown.edu/. Accessed 2014.

12. **Abstrackr**. http://abstrackr.cebm.brown.edu/. Accessed 2014

13. **Abstrackr source code** [https://github.com/bwallace/abstrackr-web]. Access date 2014

14. Wallace B, Small K, Brodley C, Lau J, Trikalinos T. **Deploying an interactive machine learning system in an evidence-based practice center**. Proc 2nd ACM SIGHIT Symp Int Heal informatics 2012.

15. Wallace B, Small K, Brodley C, Lau J, Schmid C, Bertram L, et al. **Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining**. Genet Med. 2012;14:663–9.

16. Jonnalagadda S, Petitti D. **A new iterative method to reduce workload in the systematic review process**. Int J Comput Biol Drug Des. 2013;6:5–17.

17. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. **A new algorithm for reducing the workload of experts in performing systematic reviews**. J Am Med Informatics Assoc. 2010;17:446–53.

18. Holmes M, Rathbone J, Littlewood C, Rawdin A, Stevenson M, Stevens J, et al. **Routine echocardiography in the management of stroke and transient ischaemic attack: a systematic review and economic evaluation**. Health Technol Assess (Rockv). 2014;18.

19. Papaioannou D, Rafia R, Rathbone J, Stevenson M, Buckley Woods H, Stevens J. **Rituximab for the first-line treatment of stage III–IV follicular lymphoma (review of Technology Appraisal No. 110): a systematic review and economic evaluation**. National Institute for Health Research. Health Technol Assess (Rockv). 2012;16:1–253.

20. Asano T, McLeod R. **Dietary fibre for the prevention of colorectal adenomas and carcinomas**. Cochrane Database Syst Rev. 2002;2.

21. Xiaoqing G, Tongguang N, Hongyuan W. **New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical Datasets Classification**. Sci World J. 2014;1–12.

22. Wu G, Chang E. KBA: **kernel boundary alignment considering imbalanced data distribution**. IEEE Trans Knowl Data Eng. 2005;17:786–95.

23. **Frequency of updating Cochrane Reviews** [http://www.cochrane.org/editorial-and-publishing-policy-resource/cochrane-review-updates]

24. Hopewells S, Clarke M, Moher D, Wager E, Middleton P, Altman D, et al. **CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration**. PLoS Med. 2008; 5:e20.

25. Ng L, Pitt V, Huckvale K, Clavisi O, Turner T, Gruen R, et al. **Title and Abstract Screening and Evaluation in Systematic Reviews (TASER): a pilot randomised controlled trial of title and abstract screening by medical students**. Syst Rev. 2014; 3:e1–8.

26. Small K, Wallace B, Brodley C, Trikalinos T. **The constrained weight space SVM: learning with ranked features**. In: Proc 28th Int Conf Mach Learn Bellevue, WA, USA. 2011.

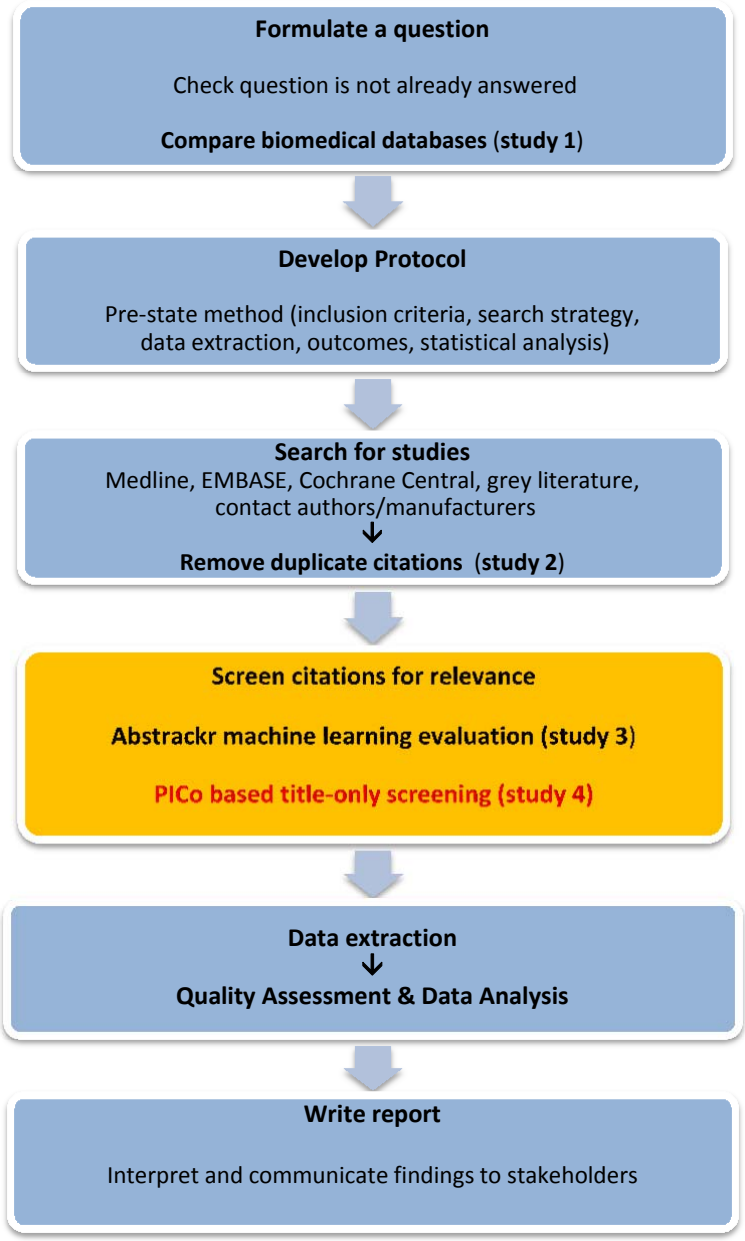
27. **Embase project** [http://www.metaxis.com/embasepublic/] presentation Washington HTAi 2014.pptx]

28. O Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. **Using text mining for study identification in systematic reviews: a systematic review of current approaches**. Syst Rev. 2015; 4:e1–22.

Copyright

Chapter 6

Screening citations using PICo based title-only screening



Key steps for conducting a systematic review and where studies for this PhD are focussed

In Chapter 5, it was demonstrated that text mining technology via machine learning is becoming increasingly feasible for systematic reviews. Chapter 6 describes a new method of screening citations automatically by utilising Boolean operator title field search methods. The methods and results are discussed along with its application within the systematic review research community.

Summary

The following paper is complete in its original aims. Post hoc tests are currently being performed to test the application of PICo based title-only screening with a machine learning algorithm to determine whether the machine learning phase can be fully automated.

6.1 Expediting citation screening using PICo based title-only screening for identifying studies in scoping searches and rapid reviews

John Rathbone, Loai Albarqouni, Mina Bakhit, Elaine Beller, Oyungerel Byambasuren, Tammy Hoffmann, Anna Mae Scott, Paul Glasziou

Abstract

Background

Citation screening for scoping searches and rapid review is time-consuming and inefficient, often requiring days or sometimes months to complete. We examined the reliability of PICo based title-only screening using keyword searches based on the PICO elements - Participants, Interventions, and Comparators, but not the Outcomes.

Methods

A convenience sample of 10 datasets, derived from the literature searches of completed systematic reviews, was used to test PICo based title-only screening. Search terms for screening were generated from the inclusion criteria of each review, specifically the PICo elements - Participants, Interventions and Comparators. Synonyms for the PICo terms were sought, including alternatives for clinical conditions, trade names of generic drugs and abbreviations for clinical conditions, interventions and comparators. The MeSH database, Wikipedia, Google searches and online thesauri were used to assist generating terms. Title-only searches were performed in Endnote X7 reference management software using OR Boolean operator. Outcome measures were recall of included studies and the reduction in screening effort. Recall is the proportion of included studies retrieved using PICo title-only screening out of the total number of included studies in the original reviews. The percentage reduction in screening effort is the proportion of records that do not need to be screened.

Results

Across the 10 reviews the reduction is screening effort ranged from 11% to 78% with a median reduction of 53%. In 9 systematic reviews, the recall of included studies was 100%. In one review (oxygen therapy), 4 of 5 reviewers missed the same included study (median recall: 67%). A post-hoc analysis was performed on the dataset with the lowest reduction in screening effort (11%), and was rescreened using only the intervention and comparator 2 keywords, and omitting keywords for participants. The reduction in screening effort increased to 57% and the recall of included studies was maintained (100%).

Conclusions

PICo based title-only screening can expedite citation screening for scoping searches and rapid reviews by reducing the number of citations needed to screen, but requires a thorough workup of the potential synonyms and alternative terms.

Introduction

There is no universal definition of what constitutes a scoping search although various criteria have been proposed [1],[2],[3]. In general, scoping searches are useful to attain a preliminary assessment of the size and scope of research literature, and to help assess the feasibility of conducting research, including determining whether clinical questions have previously been evaluated, or are up to date, and for estimating time-frames and budgetary considerations. Similarly, rapid reviews have no universally agreed upon definition but typically are a form of knowledge synthesis where some components of the systematic review process are simplified or omitted to produce information in a timely manner[4].

Scoping searches and rapid reviews both seek knowledge using a less formalised and rigorous methodology compared with systematic reviews. Rapid reviews attempt to expedite work by modifying tasks that traditional systematic reviews eschew due to the concerns over data loss[5]. Some tasks that are modified include literature searching, which may be expedited by restricting the number of databases searched[4], restricting by date range, language types[5], or omitting grey literature searches. Other strategies include restricting the number of personnel who screen studies, abstract data and assess risk of bias[4].

Citation screening of title and abstract is time-consuming because of the large number of citations typically retrieved (the average retrieval from a PubMed search produces 17,284 citations[6]) and is imprecise with often over 98% of citations from systematic searches excluded after title/abstract and full text screening([7],[8],[9],[10],[11],[12],[13],[14],[15],[16]). Titles of published studies typically incorporate the main components of a study design which can be categorised into the PICo components (Participants, Intervention, and Comparator, but not the Outcome). Therefore, screening restricted to the title field using the PICo components and the associated synonyms should identify the corpus of relevant studies whilst also being more precise, due to the constrained screening method. The aim of this study was to investigate the feasibility of conducting PICo based title-only screening primarily for scoping searches and rapid

reviews.

Methods

A convenience sample of 10 datasets derived from the literature searches of completed systematic reviews was used to test the PICo based title-only screening. Seven datasets[7],[9],[10],[11],[12],[14],[16] available to the authors were used, and an additional 3 datasets[8],[13],[15] were created by replicating the search strategy from the published reviews. These three reviews were selected prima facie based on being intervention studies that contained adequately reported search strategies and study inclusion details. We used a convenience sample of 5 reviewers, (3 clinicians, and 2 non-clinicians) based at the Centre for Research in Evidence-Based Practice, Bond University to assess the reliability and reproducibility of PICo based title-only screening. Each reviewer screened all 10 systematic reviews, and had prior knowledge of evidence-based practice and systematic review methodology.

Each reviewer independently compiled a list of search terms derived from the inclusion criteria of the reviews, specifically, the (P) Participants, (I) Interventions and (C) Comparators, but not the Outcomes. PICo synonyms including drug trade names and alternate names for clinical conditions were sought in MeSH database, Wikipedia, online thesauri and Google searches. Typically, 3-4 synonyms were generated for each term, but there was no restriction on the number of terms (see appendix 1). Keywords with both British and American spellings were used, and keywords with different suffixes were truncated using an asterisk. PICo based title-only searches were performed in Endnote X7 reference management software using ‘OR’ Boolean operator (see appendix 2).

Outcome measures (Box 1) were the recall of included studies and the reduction in screening effort (RSE). Recall is the proportion of included studies retrieved using PICo title-only screening out of the total number of included studies in the original reviews. The percentage reduction in screening effort is the proportion of records that do not need to be screened. This was reported individually for each reviewer, and as the median value across the 5 scores. A post-hoc analysis was performed

with one of the datasets (Parkinson's) to examine the impact of screening using only keywords for the (I) intervention and (C) comparator and omitting keywords for (P) participants.

Box 1.

Recall of included studies

$$recall\ (\%) = \frac{number\ of\ included\ studies\ retrieved}{total\ number\ of\ included\ studies} \times 100$$

Reduction in Screening Effort (RSE)

$$Precision = \frac{Number\ of\ included\ studies\ retrieved}{Total\ number\ of\ records\ retrieved}$$

$$Screening\ effort\ (SE) = \frac{1}{Precision}$$

$$Reduction\ in\ screening\ effort\ RSE\ (\%) = \frac{SE_{method1^{\dagger}} - SE_{method2^{\ddagger}}}{SE_{method1^{\dagger}}} \times 100$$

†method1 is current practice (screening all records).
‡method2 is PICo based title-only screening.

Results

Ten systematic reviews were evaluated with a total of 31,359 records. Reduction in screening effort across the reviews (Figure 1) ranged from 11% (Parkinson's review) to 78% (Phenytoin review) with a median reduction in screening effort of 53%. The recall of includable studies was 100% in 9 of the 10 reviews. In the oxygen therapy review, 4 of 5 reviewers missed the same included study (median recall: 67%).

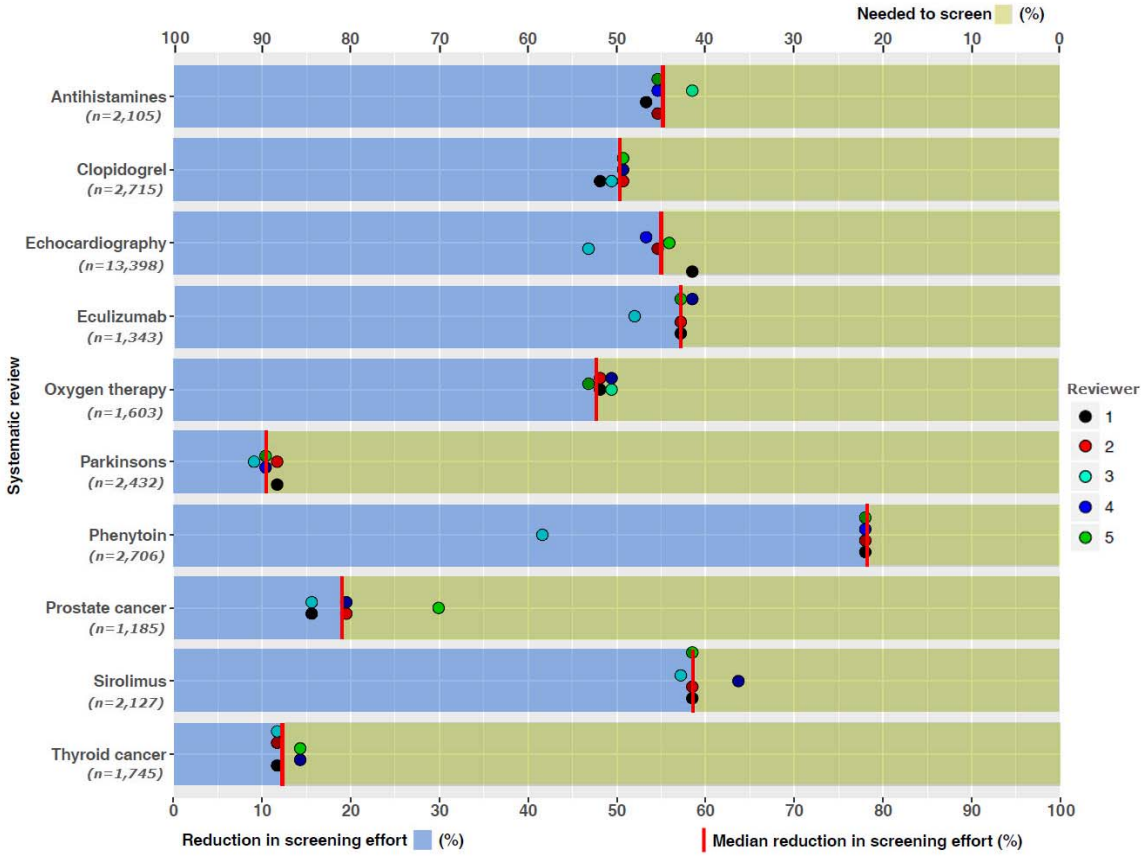


Figure 1

Summary of the median reduction (|) in screening effort, the individual reviewer reduction in screening effort (coloured dots), and the percentage of citations remaining that are needed to screen across 10 systematic reviews using PICo based title-only screening.

Post-hoc analysis

The minimal reduction in screening effort in the Parkinson's dataset was principally due to the keyword 'Parkinson' retrieving 80% of all records. A post-hoc analysis was performed to determine if complete recall could be maintained and reduction in screening effort improved when relying only on keywords for the intervention(s) and comparator(s), but not the participants. Screening without type of participants improved the median reduction in screening effort from 11% to 57%, and the recall of included studies was 100% (Figure 2).

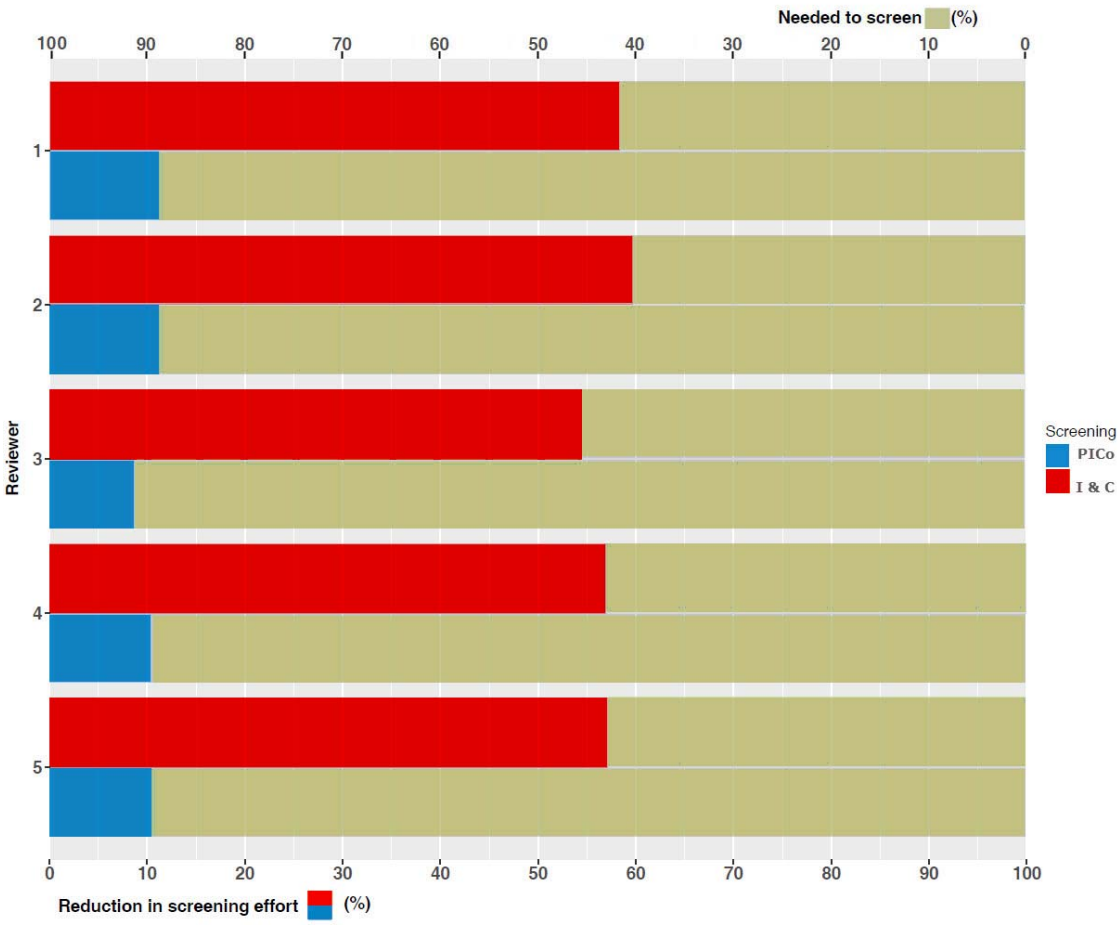


Figure 2.

Summary of the individual reviewer reduction in screening effort using PICo based title-only screening (■) and Intervention and Comparator based title-only screening (■), and the percentage of citations remaining that are needed to screen for the Parkinson's dataset.

Discussion

Our results indicate that PICo based title-only screening considerably reduces the workload of citation screening, maintains high recall of relevant studies, and can be used to expedite scoping searches and rapid reviews.

Reduction in screening effort

The reduction in screening effort ranged from 11 to 78% with 7 of the datasets having a reduction in screening effort above 50%. The two prognostic review datasets (Prostate and Thyroid cancer) had a median reduction in screening effort of 12% and 19%, however, these reviews used a more focused search and are atypical of most search strategies. The post-hoc analysis was undertaken because the reduction in screening effort was minimal in the Parkinson's dataset due to 80% of the citations containing the keyword 'parkinson' or variations e.g. 'parkinsonian' in the title field, and therefore the median reduction in screening was only 11%; the post-hoc analysis found that restricting the PICo search terms to only the intervention and comparator maintained 100% recall and improved the reduction in screening effort to 57%. This could be a useful strategy to maintain precision where a particular PICo term is overrepresented within a dataset and minimal reduction in screening effort is achievable when initially screening using all 3 PICo search terms.

The median reduction in screening effort was 53% but varied considerably from 11% to 78%. PICo based title-only screening would be of limited benefit to expedite tasks when the reduction in screening is only 10-20%, unless datasets were large (unlike the prostate and thyroid cancer datasets), but for searches that are not highly focused considerable saving can be achieved. In addition, general searches conducted in MEDLINE typically produce over 17,000 citations[44], suggesting that most searches are not highly focused and these would also benefit by applying PICo based title-only searching. Care must be taken to ensure British and American spellings and suffix variations are incorporated into the keywords screening, and that compound terms e.g. 'transoesophageal echocardiography' are entered as separate search terms to allow for variations in word order, otherwise relevant citations could potentially be missed when using PICo based title-only screening.

Recall

The recall was 100% in 9 of 10 systematic reviews. One reviewer, a clinician, identified all included studies across the 10 reviews including the oxygen therapy review; however, 4 reviewers missed the same included study in the oxygen therapy review. 'Ventilation' was used in the title as an alternative term for oxygen therapy, and this was not listed in the MeSH database, nor found whilst searching other resources, and therefore subject knowledge was needed to identify the study. Nonetheless, for other datasets PICo based title-only screening was reliable.

Strengths and limitation of the research

The strengths in this study were that 10 datasets were used to test the hypothesis that using PICo based title-only screening could retrieve all studies that should have been found and reduce the number of citations to screen. Also, the results were reproducible for recall in 9 of 10 datasets, and the methodology is simple and easily implemented by reviewers or information specialists with knowledge of screening and Endnote software. The datasets used were a convenience sample and reduction in screening effort may differ with different clinical specialities and study designs. Nonetheless, in this study, the sample of reviews tested included a variety of clinical specialties, different types of interventions and different study designs, such as diagnostic accuracy, prognostic, and intervention studies.

Applicability

The limiting factor for the applicability for screening is the presence or absence of either controlled or consistent vocabulary. The high recall and improvement in the reduction in screening effort was due to the sample datasets using clearly defined terms for (P) clinical conditions, (I) interventions and (C) comparators, but using PICo based title-only screening where the ontology is less clearly defined (e.g. where there are no MeSH terms indexed) could potentially affect recall; in such scenarios PICo based title-only screening may be unsuited for rapid review but would remain useful for scoping searches since identifying all studies is not the objective. This potential for error, however, could be allayed by including topic experts to help compile search terms. However, it has been shown that the retrieval of relevant studies for inclusion can be impaired in rapid reviews when the number

of databases searched, or the number of screeners is restricted[17]. Similarly, traditional title and abstract screening for systematic review can be imperfect with relevant studies wrongly excluded[18]. This screening methodology could also be applied to systematic review screening where one reviewer examines all records whilst a second reviewer screens the sub-set identified from PICo based-title screening.

This study has examined expediting screening on the assumption that titles of articles will include at least one of the PICo components to enable a focused title-based search to identify all relevant studies and minimise the number of citations to screen. Other methods have been developed to expedite screening using semi-automated predictive algorithms that 'learn' to distinguish relevant and irrelevant citations[19]. The recall and reduction in screening effort from PICo based title-only screening are similar to those achieved with semi-automated predictive algorithms[19],[20],[21]. However, semi-automated screening algorithms require an initial training-set (typically ~25% of the total citations) to be manually screened in order to train the algorithm. This step could be expedited by incorporating PICo based title-only screening to generate a sub-set of citations to train the algorithm and dispense with manual training. Further work is needed to explore how PICo screening can be incorporated into machine learning technologies to further accelerate the training of datasets.

Conclusion

PICo based title-only screening can expedite citation screening for scoping searches and rapid reviews by reducing the number of citations to screen, but requires a thorough workup of the potential synonyms and alternative terms.

Abbreviations:

MEDLINE: Medical Literature Analysis and Retrieval System Online

MeSH: Medical Subject Headings

PICo: Participant, Interventions, Comparators, outcomes

RSE: Reduction in Screening Effort

Acknowledgments:
Our thanks to Justin Clark and David Honeyman for replicating the search strategies, and Evelyne Rathbone for graphing the data with R statistical program.

Competing interests:
The authors declare that they have no competing interests.

Sources of funding:
NHMRC Australia Fellowship: GNT05275

Authors' contributions:
JR, EB, TH, PG contributed to the study design and concept. JR devised the testing and analysis and drafted the initial manuscript. JR, LA, MB, OB, AMS created the screening terms. All authors contributed to the manuscript and revisions, and approved the final manuscript.

Appendix 1. Example of PICo based search terms used for screening Clopidogrel and Aspirin versus Aspirin Alone for Stroke Prevention: A Meta-Analysis

Inclusion criteria:

Participants - People with stroke or transient ischaemic attack

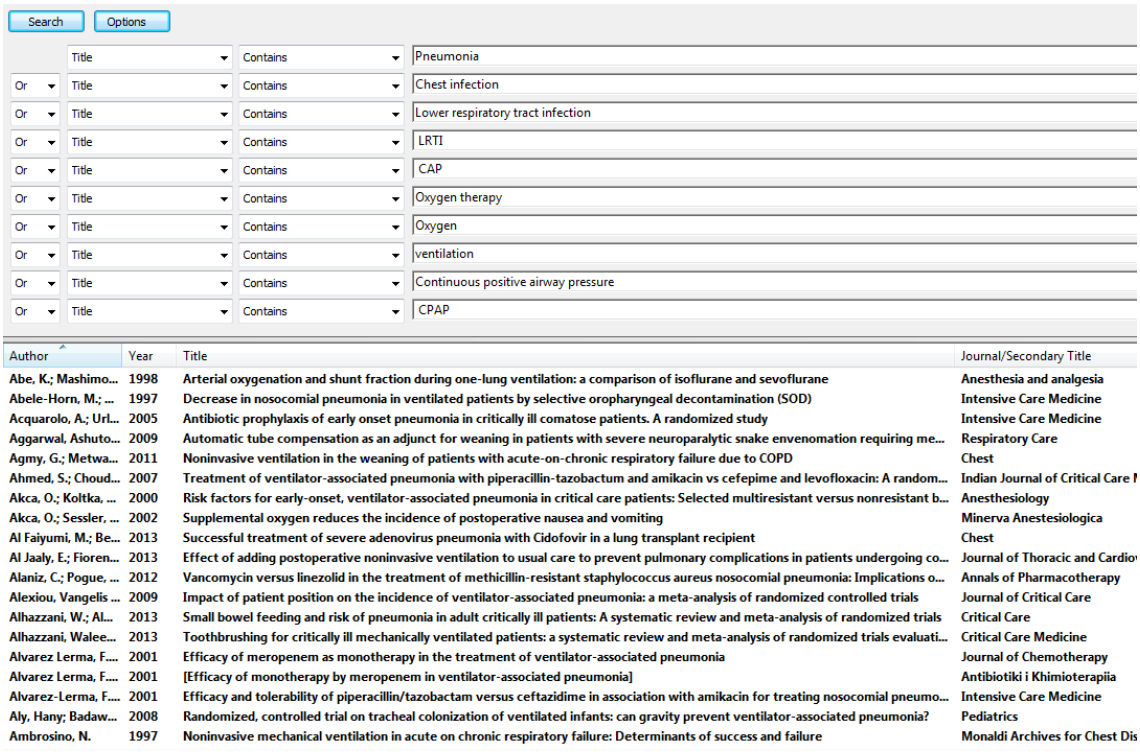
Intervention - Clopidogrel and aspirin

Comparator - Aspirin

PICo	**Alternate name**	**Alternate name**	**Alternate name**
Stroke	Intracranial Embolism and Thrombosis	Intracranial Arteriosclerosis	
Transient ischaemic attack	TIA	Brain Stem Ischemia	Transient Cerebral Ischemia
Clopidogrel	plavix	iscover	
Aspirin	Acetylsalicylic acid	ASA	2-(Acetyloxy)benzoic Acid

Appendix 2.

Example of PICo based title-only screening using OR Boolean operator in Endnote reference management software (Oxygen therapy)



Search | Options

	Title	Contains	Pneumonia
Or	Title	Contains	Chest infection
Or	Title	Contains	Lower respiratory tract infection
Or	Title	Contains	LRTI
Or	Title	Contains	CAP
Or	Title	Contains	Oxygen therapy
Or	Title	Contains	Oxygen
Or	Title	Contains	ventilation
Or	Title	Contains	Continuous positive airway pressure
Or	Title	Contains	CPAP

Author	Year	Title	Journal/Secondary Title
Abe, K.; Mashimo...	1998	Arterial oxygenation and shunt fraction during one-lung ventilation: a comparison of isoflurane and sevoflurane	Anesthesia and analgesia
Abele-Horn, M.; ...	1997	Decrease in nosocomial pneumonia in ventilated patients by selective oropharyngeal decontamination (SOD)	Intensive Care Medicine
Acquarolo, A.; Url...	2005	Antibiotic prophylaxis of early onset pneumonia in critically ill comatose patients. A randomized study	Intensive Care Medicine
Aggarwal, Ashuto...	2009	Automatic tube compensation as an adjunct for weaning in patients with severe neuroparalytic snake envenomation requiring me...	Respiratory Care
Agmy, G.; Metwa...	2011	Noninvasive ventilation in the weaning of patients with acute-on-chronic respiratory failure due to COPD	Chest
Ahmed, S.; Choud...	2007	Treatment of ventilator-associated pneumonia with piperacillin-tazobactum and amikacin vs cefepime and levofloxacin: A random...	Indian Journal of Critical Care M
Akca, O.; Koltka, ...	2000	Risk factors for early-onset, ventilator-associated pneumonia in critical care patients: Selected multiresistant versus nonresistant b...	Anesthesiology
Akca, O.; Sessler, ...	2002	Supplemental oxygen reduces the incidence of postoperative nausea and vomiting	Minerva Anestesiologica
Al Faiyumi, M.; Be...	2013	Successful treatment of severe adenovirus pneumonia with Cidofovir in a lung transplant recipient	Chest
Al Jaaly, E.; Fioren...	2013	Effect of adding postoperative noninvasive ventilation to usual care to prevent pulmonary complications in patients undergoing co...	Journal of Thoracic and Cardio
Alaniz, C.; Pogue, ...	2012	Vancomycin versus linezolid in the treatment of methicillin-resistant staphylococcus aureus nosocomial pneumonia: Implications o...	Annals of Pharmacotherapy
Alexiou, Vangelis ...	2009	Impact of patient position on the incidence of ventilator-associated pneumonia: a meta-analysis of randomized controlled trials	Journal of Critical Care
Alhazzani, W.; Al...	2013	Small bowel feeding and risk of pneumonia in adult critically ill patients: A systematic review and meta-analysis of randomized trials	Critical Care
Alhazzani, Walee...	2013	Toothbrushing for critically ill mechanically ventilated patients: a systematic review and meta-analysis of randomized trials evaluati...	Critical Care Medicine
Alvarez Lerma, F....	2001	Efficacy of meropenem as monotherapy in the treatment of ventilator-associated pneumonia	Journal of Chemotherapy
Alvarez Lerma, F....	2001	[Efficacy of monotherapy by meropenem in ventilator-associated pneumonia]	Antibiotiki i Khimioterapiia
Alvarez-Lerma, F....	2001	Efficacy and tolerability of piperacillin/tazobactam versus ceftazidime in association with amikacin for treating nosocomial pneumo...	Intensive Care Medicine
Aly, Hany; Badaw...	2008	Randomized, controlled trial on tracheal colonization of ventilated infants: can gravity prevent ventilator-associated pneumonia?	Pediatrics
Ambrosino, N.	1997	Noninvasive mechanical ventilation in acute on chronic respiratory failure: Determinants of success and failure	Monaldi Archives for Chest Dis

References

1. **The Joanna Briggs Institute Reviews' Manual 2015: Methodology for JBI Scoping Reviews** [http://joannabriggs.org/assets/docs/sumari/Reviewers-Manual_Methodology-for-JBI-Scoping-Reviews_2015_v2.pdf]

2. Levac D, Colquhoun H, O'Brien K: **Scoping studies: advancing the methodology**. Implement Sci 2010, **5**.

3. Arksey H, O'Malley L: **Scoping studies: toward s a methodological framework**. Int J Soc Res Methodol Theory Pract 2005, **8**:19–32.

4. Tricco A, Antony J, Zarin W, Strifler L, Ghassenmi M, Ivory J, Perrier L, Hutton B, Moher D, Straus S: **A scoping review of rapid review methods**. BMC Med 2015, **13**.

5. Ganann R, Ciliska D, Thomas H: **Expediting systematic reviews: methods and implications of rapid reviews**. Implement Sci 2010, **5**:56.

6. Islamaj Dogan R, Murray GC, Névéol A, Lu Z: **Understanding PubMed user search behavior through log analysis.** Database J Biol databases curation 2009, **2009**:1.

7. De Sutter A, Saraswat A, van Driel M: **Antihistamines for the common cold**. Cochrane Database Syst Rev 2015, **29**.

8. Tan S, Xiao X, Ma H, Zhang Z, Chen J, Ding L, Yu S, Xu R, Yang S, Huang X, Hong H: **Clopidogrel and Aspirin versus Aspirin Alone for Stroke Prevention: A Meta-Analysis**. PLoS One 2015, **13**.

9. Holmes M, Rathbone J, Littlewood C, Rawdin A, Stevenson M, Stevens J, Archer R, Evans P, Wang J: **Routine echocardiography in the management of stroke and transient ischaemic attack: a systematic review and economic evaluation**. Health Technol Assess 2014, **18**.

10. Rathbone J, Kaltenthaler E, Richards A, Tappenden P, Bessey A, Cantrell A: **A systematic review of eculizumab for atypical haemolytic uraemic syndrome (aHUS)**. BMJ Open 2013, **3**:1–11.

11. **Oxygen therapy for pneumonia**. (In Press).

12. Ren S, Rathbone J, Cooper K, Gomersall T, Stevens J, Harnan S, Simpson E, Sutton A, Anderson J, Cooper J, Smith H, Shaikh S: **The Efficacy and Safety of Pharmacological Therapies Used for Advanced Parkinson's Disease: A Systematic Review and Network Meta-Analysis**. (In Press).

13. Zafar S, Khan A, Ghauri A, Shamim M: **Phenytoin versus Leviteracetam for seizure** 14 **prophylaxis after brain injury - a meta analysis**. BMC Neurol 2012, **12**.

14. Bell K, Del Mar C, Wright G, Dickinson J, Glasziou P: **Prevalence of incidental prostate cancer: A systematic review of autopsy studies**. Int J Cancer 2015, **137**:1749–57.

15. Asrani S, Leise M, West C, Murad M, Pedersen R, Erwin P, Tian J, Wiesner R, Kim W: **Use of sirolimus in liver transplant recipients with renal insufficiency: a systematic review and meta-analysis**. Hepatology 2010, **52**:1360–70.

16. Furuya-Kanamori L, Bell K, Clark J, Glasziou P, Doi S: **Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis**. J Clin Oncol 2016:[Epub ahead of print].

17. Pham M, Waddell L, Rajic A, Sargeant J, Papadopoulos A, RcEven S: **Implications of applying methodological shortcuts to expedite systematic reviews: three case studies using systematic reviews from agri-food public health**. Res Synth Methods 2016, **7**:433–446.

18. Ng L, Pitt V, Huckvale K, Clavisi O, Turner T, Gruen R, Elliott J: **Title and Abstract Screening and Evaluation in Systematic Reviews (TASER): a pilot randomised controlled trial of title and abstract screening by medical students**. Syst Rev 2014, **3**:e1–8.

19. Wallace B, Small K, Brodley C, Lau J, Schmid C, Bertram L, Lill C, Cohen J, Trikalinos T: **Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining**. Genet Med 2012, **14**:663–9.

20. Rathbone J, Hoffmann T, Glasziou P: **Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers**. Syst Rev 2015, **4**:80.

21. Wallace B, Small K, Brodley C, Lau J, Trikalinos T: **Deploying an interactive machine learning system in an evidence-based practice center**. Proc 2nd ACM SIGHIT Symp Int Heal informatics 2012.

Chapter 7
Discussion

John Rathbone

7.1 Summary

This chapter briefly contextualises the development of systematic reviews and the current impasse which has provoked the need for automation technologies. It discusses the four individual studies and how they have contributed to the field of automation, summarises the thesis findings including the strengths and limitations of current automation technologies, the barriers and facilitators to the development and implementation of automation technologies, suggestions to assist with the actualization of automation technologies, and concluding remarks.

7.2 Overview of research problem

Systematic reviews developed from the need identified in the 1970s to establish a corpus of evidence by medical specialty to inform clinicians of best practice using unbiased, reliable and reproducible methods. By the 1990s this approach had begun to replace opinion-based medicine with *Evidence*-Based Medicine (which includes but is not limited to systematic reviews); however, the proponents of EBM did not foresee the growth in research and the consequent increased costs and unsustainability of research synthesis. Consequently, organisations such as the Cochrane Collaboration struggled to ensure reviews remained up to date as new trial data were published and ultimately were forced to down-grade their idealistic goals, as envisioned by Archie Cochrane. The focus, understandably for such organisations, has been to improve the quality of systematic reviews. This has led to the introduction of incremental improvements, e.g. risk of bias assessment, summary of findings tables, PRISMA flow charts and checklist. However, the validity of each review is undermined when the latest trial data are not incorporated in a timely manner, and undermines the *raison d'etre* of the Cochrane Collaboration - to produce the best available evidence.

7.3 Development of an international collaboration

More recently, there has been wider recognition of the unsustainability of research synthesis and consequently an initiative was taken in 2015 that led to the formation of the International Collaboration for the Automation of Systematic Reviews (ICASR)[45] by groups of researchers who are interested in progressing the automation of systematic reviews. The purpose of the collaboration was to discuss the development of automation technologies, and to produce policy documents which set out their aims and objectives, and to delegate tasks to groups with the appropriate expertise. This is an important step in recognition of the potential of technology to assist with the problems facing systematic reviewers. It is also the beginning of an effort to coordinate strategies to overcome these problems, which has been largely absent from the policy documents of organisations responsible for the production of systematic reviews.

It was with a similar interest to those principles established by ICASR that this PhD developed, with many of the objectives of ICASR overlapping with this body or work. Additional interest arose from personally observing and experiencing the inefficiencies and lack of progress towards better working practices for secondary health research. The different tasks required of a systematic review pose different challenges with some of those tasks being more readily automatable than others, since current technology is unable to automate all the tasks. Different approaches have been pursued by automation teams with some endeavours being partially successful but with seemingly insurmountable barriers to further progress[46], whilst others have proven to be fully automatable yet have remained as conceptual research requiring additional impetus to produce a consumer product[47]. The four projects undertaken in this PhD have contributed to our understanding of the strengths and limitations of semi-automation techniques. Furthermore, the projects have enhanced our understanding of how automation can be used for future adaptation and/or integration with other systems, or replaced with better systems as they are developed.

7.4 Comparing bibliographic databases

This study was devised to investigate the performance and reliability of biomedical databases as a resource for identifying systematic reviews. This research question and its findings were unique because no previous research had examined the performance of biomedical databases for identifying systematic reviews. The findings illustrated that EMBASE had the best sensitivity but with a trade-off with specificity and therefore greater numbers of records to screen. In contrast, the Cochrane library had a clear advantage with specificity, and therefore fewer records to screen. However, none of the databases identified all the systematic reviews and the use of search filters was a limitation affecting the sensitivity of biomedical databases.

Since the paper was published, some of the smaller bibliographic databases have expanded their content substantially due to incorporating other database provider's content e.g. PubMed records have been added to TRIP database and this may have improved the sensitivity of searches. Biomedical databases such as PubMed are

also affected by the growth in research which is impacting on the ability of PubMed to continue coding citations with medical subject headings (MeSH) which are used to improve record retrieval by indexing articles in the database with a controlled vocabulary thesaurus that enables users to search at various levels of specificity. The increasing costs are not sustainable and in the future if PubMed abandons adding MeSH terms to citations researches will face additional challenges to identify trials. Text mining will become crucial to searching for studies if MeSH terms are discontinued and this may have the unexpected benefit of spurring on further research.

7.5 Deduplication

Record deduplication has previously been overlooked as a research priority for systematic reviews and researchers have been dependent upon existing software which has barely advanced since its inception. The deduplication algorithm developed in this research project is noteworthy, having progressed from pure academic enquiry into a fully operational open access application which is now used by researchers throughout Europe, Australasia and North America. Several organisations including the UK Cochrane Collaboration, and Covidence in Melbourne, Australia are investigating integrating the deduplication application into their software systems. The National Institute for Clinical Excellence, in the UK, is proposing to develop a similar fuzzy-logic deduplication program, and is using the datasets compiled during the PhD research for their own evaluation. The deduplication research project is an example of advancing existing processes, however, scope remains to improve duplicate detection using strategies previously outlined in chapter 4. How this will be pursued in the future is unclear because of the limited resources currently committed to this field of research.

7.6 Title and abstract screening - Abstrackr

The third study was devised to investigate the feasibility and reliability of semi-automated screening. The current practice of reviewers manually screening thousands of citations can take weeks or months to complete and is one of the biggest time-consuming steps for reviewers, and therefore one of the more

important tasks to automate. The research demonstrated the potential benefits that semi-automated screening can provide with reducing screening effort by up to 80%.

The paper was cited as one of the most influential papers published by BioMed Central[43] in 2015. Since the publication, research interest in automation technologies has continued to grow with Howard (2016)[48] developing a text mining algorithm to rank citations by relevance (similar to EPPI-Reviewer[49] software). Text mining algorithms have also been developed with the aim of achieving high recall by incorporating 'voting' strategies[50] that prioritise citations that receive at least one vote. Other techniques have been explored to expedite citation screening by utilising citation networks which assume that studies meeting the inclusion criteria of a systematic review will form a network of connectivity where studies are co-cited either directly or indirectly. Belter (2015)[46] investigated this concept as an alternative literature search method for expediting the identification of studies and found that the screening effort was reduced by over 50%; however, recall was only 74%. One barrier encountered was that some studies included in the reviews (but not retrieved by citation searching) were not indexed in Web of Science. Recall could have potentially been improved by also including Scopus and Google Scholar to improve coverage; however, many of the studies were missed simply because the citation network was incomplete, and therefore citation network searching is unlikely ever to replace current information retrieval methods.

Priority should be given to further developing automation tools into consumer level applications that have demonstrated good reliability, such as machine learning citation screening algorithms because they are not hindered by the same seemingly insurmountable barriers associated with for example, citation analysis which data mines citation relationships between articles. Such tools will probably achieve greater reliability compared with human operators. Tasks requiring subjective assessment or the synthesis of complex data, may be beyond the current capabilities of automation technologies and will continue to rely on human expertise. Semi-automation tools, however, could still provide a supportive role for the synthesis of complex data. For example, extracting numerical data from graphs has been made easier with WebPlotDigitizer[51] which is a web based tool that helps convert data plots into number values.

Text mining tools are currently at an early stage of development and adoption of such technologies has been slow due to a combination of lack of awareness, reliability doubts, and compatibility issues. Many of the earlier algorithms were developed for applications where perfect recall was not the main priority[52], and developers have questioned the feasibility of whether text mining tools can ever meet the expectation of perfect recall for a systematic review[53]. Such pessimism was understandable two decades ago, when text mining applied to systematic reviews was in its infancy. Text mining algorithms are feasible when applied to systematic reviews of randomised controlled trials of drug interventions, with results typically ranging from 99% to 100% for recall because the terminology is mostly standardised. However, applying text mining to reviews with less structured vocabulary, such as in the social sciences could be more challenging. Nevertheless, it is questionable if a small loss of data would be critical to the outcome of descriptive systematic reviews which use thematic analysis. Nonetheless, relatively little research has been conducted into text mining for systematic reviews and its potential strengths and limitations remain underexplored.

7.7 PICo based title-only screening

The fourth study investigated the potential strengths and limitations of screening citations using PICo based title-only screening as an alternative to manually screening citations for rapid reviews and scoping searches. The methodology is a unique approach to screening, relying on the premise that titles will include at least one of the three main PICo terms and therefore capture relevant records using a more focused searching method. The results validated that premise and demonstrated either very high or perfect recall accompanied by about 50% reduction in screening. The reduction in the screening effort and the high recall were similar to the findings in the Abstrackr publications and suggest that this methodology can be incorporated into machine learning algorithms to replace the ‘learning’ phase with instantaneous ‘forced learning’ to train the algorithm.

7.8 Direction of future research

There are many technical and collaborative challenges confronting automation technologies. These challenges need to be overcome or new technologies risk becoming mere academic curiosity, stuck at beta development stage without progressing to a consumer product. Technical problems that developers encounter include restricted access to online biomedical databases, journals, and trial registries due to commercial restrictions or paywalls[54]. Also, many database providers are unwilling to provide access to application programming interface (API) keys which allows unrelated software programs to communicate with one another. Objections to accessing API keys include citing copyright restriction (this was encountered during research in Chapter 3), even though the information is in the public domain, albeit in a format that prevents the automation of tasks that are currently performed manually such as citation analysis or citation enrichment. To overcome these barriers a protocol needs to be developed amongst the stake holders similar to those developed by DICOM (Digital Imaging and Communications in Medicine) which is a standard for distributing and viewing any type of medical image[55,56].

Unrestricted full text access could facilitate, for example, the use of screen scraping tools to replace manual data entry of citation details. Other issues include lack of agreement on technical standards preventing integration of different programming platforms which need to be overcome with an agreed standard to allow 'plug and play' systems. There are different strategies to automate systematic reviews and some will be more challenging because of technical barriers e.g. fully automating data extraction due to the multitude of different ways numerical results can be reported (in tables, graphs, as proportions, dichotomised data as improved/not improved). Also studies often classify patients differently when describing severity of illness, or provide insufficient information and would therefore thwart automation. As processes become increasingly automated fewer gains will be achieved with diminishing returns for the investment in development time. Collaborative barriers will prevent automation due to organisations wishing to either dominate a service or maintain current market position due to commercial interests. Some automation processes will save time and also improve accuracy such as automated results

writing tools which have the potential to reduce human error but more importantly provide an idealised text that is more meaningful and comprehensible to readers. Other automation tasks will transform current practice and substantially save time such as automated screening tools which can reduce the screening content by as much as 80% and save weeks of work. Some of the tasks that are either in current development or would benefit from automation include:

1. Replace tasks that are currently manually performed such as citation screening and data abstraction with machine learning algorithms.

2. Improve existing but imperfect semi-automation methods, e.g. duplicate detection in reference management software packages such as EndNote™ with more accurate systems such as SRA-deduplication tool, and strive for fully automated duplicate detection.

3. Continually improve existing screening applications such as Covidence, Rayyan and DistillerSR so that new applications can be easily incorporated once they become available, enabling apps to be downloaded and installed. This would overcome some current limitations that require data to be exported and reformatted from one program to another.

4. Expedite and improve the accuracy of the reporting of systematic reviews, e.g. improving the protocol, background, methods, results and discussion sections by pre-populating with a selection of structured sentences to improve comprehension. Automate writing of results using the pre-populated structured data within statistical programs to generate text describing the direction and size of treatment effects.

5. Prioritise the automation of tasks that are time-intensive e.g. citation screening, data abstraction. Prioritise automation tools that have demonstrated proof of concept but have not been further developed into a consumer level product.

Many of these research priorities are at different stages of development. Some have been investigated and have shown initial promise, such as predictive screening

tools like Abstrackr[47], or visual text mining techniques[57] but have not progressed to consumer level applications, whilst other applications are in the early stage of development such as RevMan Replicant which will eventually populate the results section of a systematic review with a first draft describing the size, direction, and significance level of the treatment effect. Another example, is a machine learning program called RobotReviewer[35] that has recently been developed to expedite the assessment of risk of bias in Cochrane and non-Cochrane systematic reviews. The accuracy is slightly lower (~<10%) compared with human screening decisions; however, because the algorithm highlights relevant text it can assist reviewers to find information when used as a companion tool, or as an algorithmic second reviewer.

Research priorities will vary according to organisational needs and other automation tasks may take precedence. Future decisions on how best to prioritise research need to consider the current and future state of technology which may either restrict or enable automation. For example, future developments such as record linkage across biomedical databases could foreseeably enable citation enrichment and therefore supersede the need for additional research into duplicate detection. Also, workflows may change as automation processes develop. For example, PICo based title-only screening could be used to conduct an initial search across biomedical databases, followed by a machine learning algorithm that incorporates this data to complete the fully automated search, including citation analysis and future periodic update searches. As machine learning improves, the excluded studies in published systematic reviews could also be used to provide a feedback loop to enhance machine learning algorithms.

Commercial and non-commercial groups are independently developing software applications to facilitate automation. Most products available as consumer level products have been developed to assist with the screening, organisation and cataloguing of records within systematic reviews. For example, screening tools have improved the visual experience of selecting studies and assist with the tracking of conflicting screener decisions. The main benefits of these tools are primarily to improve the organisation of data, reduce human error, and improve the user experience above expediting tasks, although inevitably these improvements can

save time. The Cochrane group's own software, RevMan, automatically calculates summary statistics from structured data, generates forest plots and calculates statistical bias. Most of these developments have occurred independently and integrating automation tools into different software has been challenging because of different programming languages. Greater co-operation is needed to have a set of agreed standards to facilitate automation research including open source code and a mechanism for different software to communicate with one another by way of a standard application programming interface key, or a platform enabled to integrate different applications with 'plug and play' standardisation.

The limited replication of validating automation technologies, e.g. predictive citation screening makes it difficult to assess its applicability in other healthcare specialties where the ontology is less clearly defined. Independent evaluation of automation technologies is needed to validate initial findings and to test how well technologies perform when applied in different contexts with different datasets. By necessity, most technologies are tested on small datasets due to limited resources because validation is often the biggest development cost for automation technologies. Attempts to overcome some of these problems have been proposed by ICASR[45]. The working group developed a set of core principles (The Vienna Principles – see Box 1) which includes encouraging collaboration between automation research groups. Each automation group has different types of expertise and by collaborating and sharing ideas barriers to progress can be more readily overcome.

Box 1. The Vienna Principles

1. Systematic reviews involve multiple tasks, each with different issues, but all must be improved.
2. Automation may assist with all tasks, from scoping reviews to identifying research gaps as well protocol development to writing and dissemination of the review.
3. The processes for each task can and should be continuously improved, to be more efficient and more accurate.
4. Automation can and should facilitate the production of systematic reviews that adhere to high standards for the reporting, conduct and updating of rigorous reviews.
5. Developments should also provide for flexibility in combining and using, e.g. subdividing or merging steps and allow for different users to use different interfaces.
6. Different groups with different expertise are working on different parts of the problem; to improve reviews as a whole will require collaboration between these groups.
7. Every automation technique should be shared, preferably by making code, evaluation data and corpora available for free.
8. All automation techniques and tools should be evaluated using a recommended and replicable method with results and data reported.

7.9 Barriers and facilitators to adopting automation technologies

7.9.1 Barriers

How quickly researchers and organisations adopt semi-automation technologies is speculative. Some of the barriers may be psychological, especially if organisations are averse to risk e.g. concerning data loss, or reluctant to incorporate new technology, especially larger organisations that may be less flexible. For example, web-based screening tools are available that visually enhance screening and track and alert reviewers of conflicting decisions, but many organisations continue to use older working practices. Financial barriers may exist as some programs are fee based, but these costs would be off-set by greater efficiency and by eliminating mundane tasks. Free software that are at the beta-developmental stage will not have extensive technical support and users may not persist if technical problems are encountered.

Systematic review automation is not restricted by government regulations, safety standard requirements, or large scale development costs such as occurs in the aviation industry[58]. Rather, concerns over the accuracy and reliability will be the main barrier to accepting automation technologies even though existing processes are fallible due to human or technological errors. Algorithms are more reliable when performing tasks such as citation screening and results writing because the information uses controlled vocabulary, but becomes less reliable when using data that is less well-structured, such as the assessment of risk of bias, or extracting outcome data. However, this should not be a barrier to integrating these technologies because they could provide a complementary role, but would require co-validation of the machine learning decisions by reviewers.

Often barriers to adopting more efficient working practices are less obvious when viewed externally. For example, within the Cochrane collaboration study-based registers were developed in the 1990s, but only implemented by 12 Cochrane groups by 2005[59] even though the long-term benefits outweighed the initial set-up time and costs. Study-based registers produce highly specific search results, free of duplicates, with secondary publications linked to primary studies. Financial constraints may have prevented some groups from establishing a study-based register; however, some groups may have been less inclined to adopt study-based

registers because the main beneficiaries were the academic researchers working voluntarily for the Cochrane groups who bore the burden of these inefficiencies. In a report commissioned by the UK Department of Health[25], on the future investment in UK Cochrane groups, it was noted that some groups had a number of resources, generated from National Institute for Health Research funding, including specialist registers and recommended that Cochrane investigate how these resources can be shared.

7.9.2 Facilitators

Facilitators to increase in the adoption of automation tools include rigorous evaluation and dissemination of the benefits of these technologies through journal publications, scientific conferences and social media. The integration of technologies by large organisations such as Clarivate Analytics, Mendeley, the Cochrane Collaboration, and the National Institute for Health and Care Excellence would greatly increase the awareness and adoption of automation technologies. Large organisations are also more readily able to offer technical support and provide resources for ongoing development and therefore users will be more confident to transition to new software and working practices.

7.10 Systematic reviews as a marketing tool

Systematic reviews have contributed to our understanding of the benefits and harms of treatment, and there have been many noteworthy reviews that have changed healthcare dramatically including identifying the harms of radiotherapy in early stage breast cancer[11], identifying the benefits of corticosteroids to increase survival rates in preterm pregnancies[60], and identifying the benefits of streptokinase to reduce death in acute myocardial infarction[61]. Preventative medicine has also benefited from systematic review methodology by identifying the increased risk of lung cancer from 'second-hand' cigarette smoke exposure[62]. However, systematic reviews are not without criticism from the advocates of evidence-based medicine.

Some proponents of evidence-based medicine have argued that evidence-based medicine has become a marketing tool for the pharmaceutical industry[63], and has

been hijacked so that clinical medicine has been transformed into finance-based medicine[64]. Is it possible that systematic reviews rather than being a stalwart for evidence-based medicine are providing a 'rubber stamp' of approval for the pharmaceutical industry? Such criticism is not without supporting evidence and there are many examples where research findings were found to be misleading.

Studies have shown a correlation between financial interest and positive outcomes. For example, a comparison of 319 trials examining industry-sponsored and non-industry sponsored trials found that industry sponsored trials tended to yield favourable results for the experimental treatment[65]. It is known that industry sponsored trials typically avoid comparing their own drugs against competitors' products[66], and industry-sponsored trials are more likely than other trials to conclude that a drug is safe[67]. Our understanding of the effectiveness of drugs can be overturned when new data from previously withheld trials emerge. An example is the antiviral drug Tamiflu, which was thought to reduce the risk of pneumonia and death in patients infected with the influenza virus when the evidence for its benefits were first published. However, not all the studies had been published and following a lengthy wrangle to acquire unpublished data the updated review found that Tamiflu had limited effect on the complications of influenza[22,68]. Similarly, the selective serotonin reuptake inhibitor class of antidepressants have been shown to increase the risk of suicide in children[69]. However, the original studies did not reveal this problem because either the trial data remained unpublished or patients attempting suicide were misleadingly categorised as 'overdoses'.

It is not clear if these are isolated problems or are emblematic of a wider problem in evidence-based medicine. Identifying these problems can be difficult due to the hidden nature of missing data. Contacting regulatory agencies to acquire clinical study reports is a lengthy procedure, which sometimes can only be obtained through persistence and use of the Freedom of Information Act. The reports are often lengthy 8000 plus page documents[70] that are often incomplete and cannot be easily text-searched because they are provided as an image file. Moreover, using the Freedom of Information Act to obtain clinical study reports held by regulatory bodies is time-consuming and can take months and sometimes years to gather data[22], often in a piecemeal fashion. Searching for this type of missing data goes beyond the standard procedures recommended by the Cochrane Collaboration and is not

routinely undertaken. This is understandable given both the time required and the monetary constraints that review teams will encounter. These issues are highly pertinent to the automation of systematic reviews. If the many steps of systematic reviewing could be automated this would enable reviewers to make better use of resources pursuing and analysing clinical study reports.

7.11 Conclusions

The growth of research will continue to delay the production of systematic reviews and thus incentivise the development and adoption of automation technologies. Research teams have demonstrated an enthusiasm to begin developing automation technologies, including prototype software that is slowly being advanced into consumer level products. The pace of progress may be hindered by several factors including the reluctance of funding bodies, responsible for the commissioning of systematic reviews, to invest resources to support the development of automation technologies, and the absence of collaboration between research teams, commissioning bodies, and data repository providers to pursue common goals and adopt common standards to facilitate automation.

The research projects conducted, as part of this PhD thesis, have achieved a greater understanding of automation processes for systematic reviews. This includes the development of a consumer level deduplication product which has surpassed current practice, the evaluation and demonstration of the potential benefits of semi-automation citation screening, surveying the strengths and weaknesses of bibliographic databases to identify published systematic reviews, and developing proof of concept research into expediting the screening of citations using PICo based title-only screening.

Pursuing strategies to overcome technical barriers and develop proof of concept technologies into consumer level products, and replacing manual tasks with semi-automation, or replacing semi-automation with full automation will expedite research and save resources. Importantly, it will also allow investigators to make better use of their time to contextualise and interpret the research findings in relation to current practice, and to devote time to investigate the completeness of the evidence through

the acquisition and analysis of clinical study reports for the welfare of patients, evidence-based medicine and society.

References

1. Tricco, A. *et al.* A scoping review of rapid review methods. *BMC Med.* **13,** (2015).

2. Rossi, C. & Russo, F. in *Ancient Engineers' Inventions: precursors of the present* 235–247 (2009).

3. Murray, C. & Chu, A. The flying sidekick traveling salesman problem: Optimization of drone-assisted parcel delivery. *Transp. Res. Part C Emerg. Technol.* 86–109 (2015).

4. Olson, E. Will songs be written about autonomous cars? The implications of self-driving vehicle technology on consumer brand equity and relationships. *Int. J. Technol. Mark.* **21,** (2017).

5. Marchevsky, A., Walts, A. & MR, W. Evidence-based pathology in its second decade: toward probabilistic cognitive computing. *Hum. Pathol.* **61,** 1–8 (2017).

6. Miller, A. The future of health care could be elementary with Watson. *Can. Med. Assoc. J.* **185,** E367–E368 (2013).

7. Cochrane, A. *Effectiveness & Efficiency: Random Reflections on Health Services*. (1971).

8. Ioannidis, J. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294,** 218–288 (2005).

9. Antman, E., Lau, J., Kupelnick, B., Mosteller, F. & Chalmers, T. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* **268,** 240–248 (1992).

10. A Framework for NHMRC Assessment and Funding of Clinical Trials and Cohort Studies. (2017). at <https://aamri.org.au/wp-content/uploads/2017/07/AAMRI-submission-to-NHMRC-consultation-on-assessment-and-funding-of-clinical-trials-and-cohort-studies.pdf>

11. Stjernswärd, J. Decreased survival related to irradiation postoperatively in

early breast cancer. *Lancet* **304,** 1285–1286 (1974).

12. Gilbert, R., Salanti, G., Harden, M. & See, S. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int. J. Epidemiol.* **34,** 874–87 (2005).

13. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, Howells DW, Ioannidis JP, O. S. How to increase value and reduce waste when research priorities are set. *Lancet* **383,** 156–165 (2014).

14. Editors: Julian PT Higgins and Sally Green. *Cochrane Handbook for Systematic Reviews of Interventions*. (The Cochrane Collaboration, 2011). at <www.cochrane-handbook.org>

15. Mickenautsch, S. Research gaps identified during systematic reviews of clinical trials: glass-ionomer cements. *BMC Oral Health* (2012). doi:10.1186/1472-6831-12-18

16. Thomson, H., Thomas, S., Sellstrom, E. & Petticrew, M. *Housing Improvements for Health and Associated Socio-economic Outcomes: A Systematic Review*. (2013). at <http://www.campbellcollaboration.org/lib/project/61/>

17. Oliver, R., Reschly, D. & Wehby, J. *The Effects of Teachers' Classroom Management Practices on Disruptive, or Aggressive Student Behavior: A Systematic Review*. (2011). at <http://www.campbellcollaboration.org/lib/project/164/>

18. Villettaz, P., Gilliéron, G. & Killias, M. The Effects on Re-offending of Custodial vs. Non-custodial Sanctions: An Updated Systematic Review of the State of Knowledge. *The Cambell Collaboration* (2015). at <http://www.campbellcollaboration.org/lib/project/22/>

19. Ganann, R., Ciliska, D. & Thomas, H. Expediting systematic reviews: methods and implications of rapid reviews. *Implement Sci* **5,** 56 (2010).

20. Measuring the performance of The Cochrane Library: The Cochrane Library Oversight Committee. (2012). at <http://www.thecochranelibrary.com/details/editorial/3620281/Measuring-the-performance-of-The-Cochrane-Library.html>

21. Mills, E., Thorlund, K. & Ioannidis, J. Demystifying trial networks and network

meta-analysis. *BMJ* **14,** f2914 (2013).

22. Doshi, P., Jefferson, T. & Del Mar, C. The Imperative to Share Clinical Study Reports: Recommendations from the Tamiflu Experience. *PLoS Med* **9,** (2012).

23. Bastian, H., Glasziou, P. & Chalmers, I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *PLoS Med* **7,** e1000326 (2010).

24. Ioannidis, J. P., Chang, C. Q., Lam, T. K., Schully, S. D. & Khoury, M. J. The geometric increase in meta-analyses from China in the genomic era. *PLoS One* **8,** e65602 (2013).

25. Kleijnen J, Alderson P, Aubin J, Cairns J, Crowe S, G. P. *Evaluation of NIHR investment in Cochrane infrastructure and systematic reviews*. (2017). at <https://www.journalslibrary.nihr.ac.uk/downloads/other-nihr-research/evaluation-of-NIHR-investment-in-cochrane/NIHR_Cochrane_Report_Feb_17.pdf>

26. Editorial. The cost of salami slicing. *Nat. Mater.* **4,** 1 (2005).

27. Jackson, D., Walter, G., Daly, J. & Cleary, M. Multiple outputs from single studies: acceptable division of findings vs. 'salami' slicing. *J Clin Nurs* **23,** 1–2 (2014).

28. McGauran N, W. B. & Kreis J, Schüler YB, Kölsch H, K. T. Reporting bias in medical research - a narrative review. *Trials* **11,** 11–37 (2010).

29. Duggan, L., Fenton, M., Rathbone, J., Dardennes, R. & El-Dosoky, A. Olanzapine for schizophrenia. *Cochrane Database of Systematic Reviews* (2014). at <http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD001359.pub2/references>

30. Takwoingi, Y., Hopewell, S., Tovey, D. & Sutton, A. A multicomponent decision tool for prioritising the updating of systematic reviews. *BMJ* **347,** 7191 (2013).

31. MacLehose H, Hilton J, Mehta M, Urquhart B, Royle E, Bell-Syer S, Dooley L, S. A. Cochrane Editorial Unit report: Editorial Policy and Publishing team. at <http://community.cochrane.org/sites/default/files/uploads/inline-files/Policy and Publishing.pdf>

32. Beller EM, Chen JK, Wang UL, G. P. Are systematic reviews up-to-date at the time of publication? *Syst. Rev.* **2,** (2013).

33. Garritty, C., Tsertsvadze, A., Tricco, A. C., Sampson, M. & Moher, D. Updating systematic reviews: an international survey. *PLoS One* **5,** e9914 (2010).

34. Kiritchenko, S., de Bruijn, B., Carini, S., Martin, J. & Sim, I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inf. Decis Mak* **28,** 56 (2010).

35. Marshall, I., Kuiper, J. & BC, W. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J. Am. Informatics Assoc.* **23,** 193–201 (2016).

36. Covidence. (2017). at <https://www.covidence.org/>

37. Distiller. (2017). at <https://distillercer.com/>

38. Rayyan. (2017). at <rayyan.qcri.org>

39. Turner, S., Adams, N., Cook, A., Price, A. & Milne, R. Potential benefits of using a toolkit developed to aid in the adaptation of HTA reports: a case study considering positron emission tomography (PET) and Hodgkin's disease. *Heal. Res Policy Syst* **26,** 16 (2010).

40. Noel-Storr, A. The Trial Blazers study: crowdsourcing and Cochrane. (2014). at <http://www.cochrane.org/news/blog/trial-blazers-study-crowdsourcing-and-cochrane>

41. Webster, A., Heslop, L., Chapman, J. & Craig, J. The prevalence and impact of overt and covert duplicate publications of randomized trials in renal transplantation. in *XI Cochrane Colloquium: Evidence, Health Care and Culture; 2003 Oct 26-31; Barcelona, Spain*

42. Rathbone, J., Carter, M., Hoffmann, T. & Glasziou, P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. *Syst. Rev.* **4,** 6 (2015).

43. BioMed Central's Most Influential Systematic Review Articles. (2016). at <http://blogs.welch.jhmi.edu/WelchBlog/content/biomed-centrals-most-

influential-systematic-review-articles>

44. Islamaj Dogan, R., Murray, G. C., Névéol, A. & Lu, Z. Understanding PubMed user search behavior through log analysis. *Database J. Biol. databases curation* **2009,** 1 (2009).

45. The Vienna Principles. *Evidence-Based Research Network* (2017). at <http://ebrnetwork.org/the-vienna-principles/>

46. Belter, C. Citation analysis as a literature search method for systematic reviews. *J. Assoc. Infomation Sci. Technol.* **67,** 2766–2777 (2015).

47. Wallace, B., Small, K., Brodley, C., Lau, J. & Trikalinos, T. Deploying an interactive machine learning system in an evidence-based practice center. *Proc. 2nd ACM SIGHIT Symp. Int. Heal. informatics* (2012).

48. Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shar MR. SWIFT-Review: a text-mining workbench for systematic review. *Syst. Rev.* **5,** (2016).

49. EPPI-Reviewer 4. (2017). at <http://eppi.ioe.ac.uk/CMS/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4&>

50. Wallace, B., Trikalinos, T., Lau, J., Brodley, C. & Schmid, C. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* **11,** (2010).

51. Rohatgi, A. WebPlotDigitizer. (2017). at <http://arohatgi.info/WebPlotDigitizer>

52. Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O. & O'Blenis, P. A new algorithm for reducing the workload of experts in performing systematic reviews. *Jounal Am. Med. Informatics Assoc.* **17,** 446–53 (2010).

53. Miroslav, K. & Matwin, S. Addressing the curse of imbalanced training sets: one-sided selection. in *Proceedings of the Fourteenth International Conference on Machine Learning* (1997).

54. Carter, M. Personal communication. (2017).

55. Pianykh, O. *Digital Imaging and Communications in Medicine (Dicom) : A Practical Introduction and Survival Guide*. (Springer, 2008).

56. Drnasin, I., Grgic, M. & Gogic, G. JavaScript Access to DICOM Network and Objects in Web Browser. *J. Digit. Imaging* **10,** (2017).

57. Felizardo, K., Andery, G., Paulovich, F., Minghim, R. & Maldonado, J. A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Inf. Softw. Technol.* **54,** 1079–91 (2012).

58. Warchocki, T. in *Autonomy Research for Civil Aviation: Toward a New Era of Flight* (The National Academies of Sciences Engineering Medicine, 2014).

59. Wright, J. Study-based registers. in (2005). at <http://www.academia.edu/23517403/Study-based_registers>

60. Crowley, P., Chalmers, I. & Keirse, M. The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. *Br. J. Obs. Gynaecol.* **97,** 11–25 (1990).

61. Yusuf S, Collins R, Peto R, Furberg C, Stampfer MJ, Goldhaber SZ, H. C. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: Overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur. Heart J.* **6,** 556–85 (1985).

62. Taylor, R., Najafi, F. & Dobson, A. Meta-analysis of studies of passive smoking and lung cancer: Effects of study type and continent. *Int. J. Epidemiol.* **36,** 1048–1059 (2007).

63. Healy, D. Time to abandon evidence based medicine? *Cardiff University School of Psychology* (2012). at <http://davidhealy.org/>

64. Ioannidis, J. Evidence-based medicine has been hijacked: a report to David Sackett. *J. Clin. Epidemiol.* **73,** 82–6 (2016).

65. Flacco ME, Manzoli L, Boccia S, Capasso L, Aleksovska K, Rosso A, Scaioli G, De Vito C, Siliquini R, V. P. & JP, I. Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. *J. Clin. Epidemiol.* **68,** 811–20 (2015).

66. Lathyris, D., Patsopoulos, N., Salanti, G. & Ioannidis, J. Industry sponsorship and selection of comparators in randomized clinical trials. *Eur. J. Clin. Invest.* **40,** 1–11 (2009).

67. Golder, S. & Loke, Y. Is there evidence for biased reporting of published

adverse effects data in pharmaceutical industry-funded studies? *Br. J. Clin. Pharmacol.* **66,** 767–773 (2008).

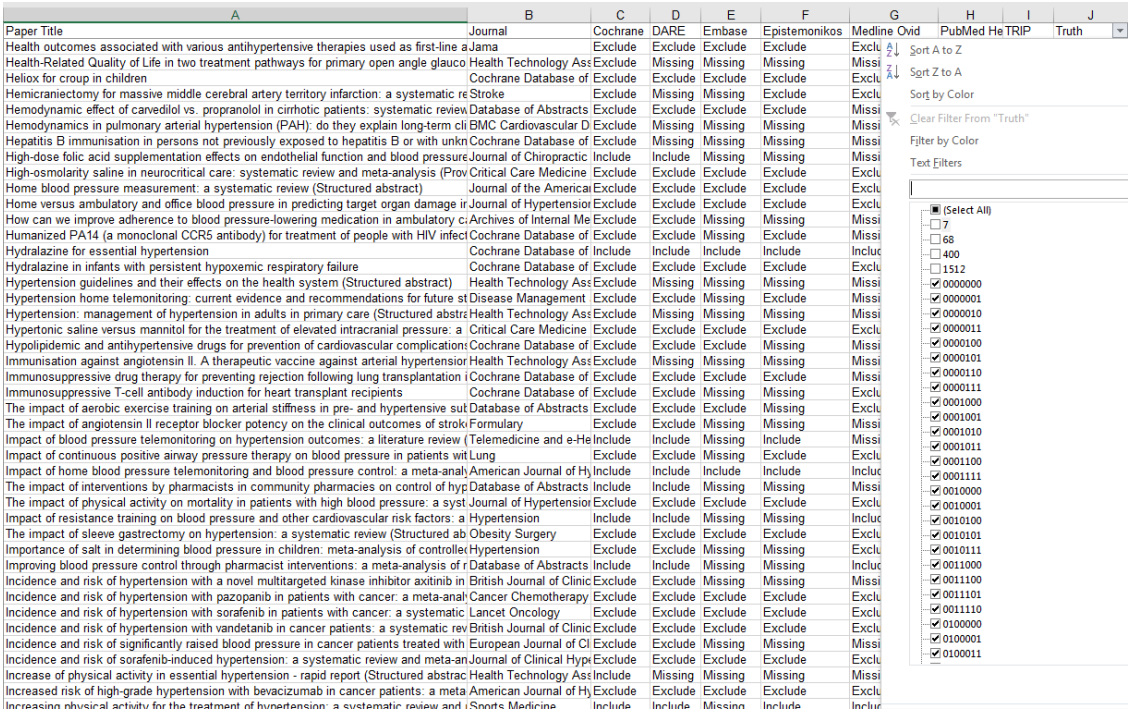
68. Jefferson T, Jones MA, Doshi P, Del Mar CB, Hama R, Thompson MJ, Spencer EA, Onakpoya I, Mahtani KR, Nunan D, Howick J, H. C. Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. *Cochrane Database Syst. Rev.* (2014). at <http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008965.pub4/abstract>

69. Sharma, T., Guski, L., Freund, N. & Gøtzsche, P. Suicidality and aggression during antidepressant treatment: systematic review and meta-analyses based on clinical study reports. *BMJ* **352,** (2016).

70. Schroll, J., Penninga, E. & Gøtzsche, P. Assessment of Adverse Events in Protocols, Clinical Study Reports, and Published Papers of Trials of Orlistat: A Document Analysis. *PLOS Med.* **13,** e1002101 (2016).

Supplementary appendix A Identifying reviews

Example of citations listed in each database and cross-compared with Truth-table created in Excel spreadsheet



Paper Title	Journal	Cochrane	DARE	Embase	Epistemonikos	Medline Ovid	PubMed He	TRIP	Truth
Health outcomes associated with various antihypertensive therapies used as first-line a	Jama	Exclude	Exclude	Exclude	Exclude	Exclu			
Health-Related Quality of Life in two treatment pathways for primary open angle glauco	Health Technology Ass	Exclude	Missing	Missing	Missing	Missi			
Heliox for croup in children	Cochrane Database of	Exclude	Exclude	Exclude	Exclude	Exclu			
Hemicraniectomy for massive middle cerebral artery territory infarction: a systematic re	Stroke	Exclude	Missing	Missing	Exclude	Exclu			
Hemodynamic effect of carvedilol vs. propranolol in cirrhotic patients: systematic review	Database of Abstracts	Exclude	Exclude	Exclude	Exclude	Missi			
Hemodynamics in pulmonary arterial hypertension (PAH): do they explain long-term cli	BMC Cardiovascular D	Exclude	Missing	Missing	Missing	Missi			
Hepatitis B immunisation in persons not previously exposed to hepatitis B or with unkn	Cochrane Database of	Exclude	Missing	Missing	Missing	Missi			
High-dose folic acid supplementation effects on endothelial function and blood pressure	Journal of Chiropractic	Include	Include	Missing	Missing	Missi			
High-osmolarity saline in neurocritical care: systematic review and meta-analysis (Prov	Critical Care Medicine	Exclude	Exclude	Exclude	Exclude	Exclu			
Home blood pressure measurement: a systematic review (Structured abstract)	Journal of the America	Exclude	Exclude	Exclude	Exclude	Exclu			
Home versus ambulatory and office blood pressure in predicting target organ damage ir	Journal of Hypertensior	Exclude	Exclude	Exclude	Exclude	Exclu			
How can we improve adherence to blood pressure-lowering medication in ambulatory c	Archives of Internal Me	Exclude	Exclude	Missing	Missing	Missi			
Humanized PA14 (a monoclonal CCR5 antibody) for treatment of people with HIV infect	Cochrane Database of	Exclude	Exclude	Missing	Exclude	Missi			
Hydralazine for essential hypertension	Cochrane Database of	Include	Include	Include	Include	Includ			
Hydralazine in infants with persistent hypoxemic respiratory failure	Cochrane Database of	Exclude	Exclude	Exclude	Exclude	Exclu			
Hypertension guidelines and their effects on the health system (Structured abstract)	Health Technology Ass	Exclude	Missing	Missing	Missing	Missi			
Hypertension home telemonitoring: current evidence and recommendations for future st	Disease Management	Exclude	Exclude	Missing	Exclude	Missi			
Hypertension: management of hypertension in adults in primary care (Structured abstra	Health Technology Ass	Exclude	Missing	Missing	Missing	Missi			
Hypertonic saline versus mannitol for the treatment of elevated intracranial pressure: a	Critical Care Medicine	Exclude	Exclude	Exclude	Exclude	Exclu			
Hypolipidemic and antihypertensive drugs for prevention of cardiovascular complication	Cochrane Database of	Exclude	Exclude	Exclude	Missing	Missi			
Immunisation against angiotensin II. A therapeutic vaccine against arterial hypertensior	Health Technology Ass	Exclude	Missing	Missing	Missing	Missi			
Immunosuppressive drug therapy for preventing rejection following lung transplantation i	Cochrane Database of	Exclude	Exclude	Exclude	Exclude	Missi			
Immunosuppressive T-cell antibody induction for heart transplant recipients	Cochrane Database of	Exclude	Exclude	Missing	Missing	Exclu			
The impact of aerobic exercise training on arterial stiffness in pre- and hypertensive sut	Database of Abstracts	Exclude	Exclude	Exclude	Missing	Exclu			
The impact of angiotensin II receptor blocker potency on the clinical outcomes of stroke	Formulary	Exclude	Exclude	Missing	Missing	Missi			
Impact of blood pressure telemonitoring on hypertension outcomes: a literature review (	Telemedicine and e-He	Include	Include	Missing	Include	Missi			
Impact of continuous positive airway pressure therapy on blood pressure in patients wit	Lung	Exclude	Exclude	Missing	Exclude	Exclu			
Impact of home blood pressure telemonitoring and blood pressure control: a meta-analy	American Journal of Hy	Include	Include	Include	Include	Includ			
The impact of interventions by pharmacists in community pharmacies on control of hyp	Database of Abstracts	Include	Include	Missing	Missing	Missi			
The impact of physical activity on mortality in patients with high blood pressure: a syst	Journal of Hypertensior	Exclude	Exclude	Exclude	Exclude	Exclu			
Impact of resistance training on blood pressure and other cardiovascular risk factors: a	Hypertension	Include	Include	Missing	Missing	Includ			
The impact of sleeve gastrectomy on hypertension: a systematic review (Structured ab	Obesity Surgery	Exclude	Exclude	Exclude	Exclude	Exclu			
Importance of salt in determining blood pressure in children: meta-analysis of controlle	Hypertension	Exclude	Exclude	Missing	Missing	Exclu			
Improving blood pressure control through pharmacist interventions: a meta-analysis of r	Database of Abstracts	Include	Include	Missing	Missing	Includ			
Incidence and risk of hypertension with a novel multitargeted kinase inhibitor axitinib in	British Journal of Clinic	Exclude	Exclude	Missing	Missing	Missi			
Incidence and risk of hypertension with pazopanib in patients with cancer: a meta-anal	Cancer Chemotherapy	Exclude	Exclude	Exclude	Exclude	Exclu			
Incidence and risk of hypertension with sorafenib in patients with cancer: a systematic	Lancet Oncology	Exclude	Exclude	Exclude	Exclude	Exclu			
Incidence and risk of hypertension with vandetanib in cancer patients: a systematic rev	British Journal of Clinic	Exclude	Exclude	Exclude	Exclude	Exclu			
Incidence and risk of significantly raised blood pressure in cancer patients treated with	European Journal of Cl	Exclude	Exclude	Missing	Missing	Missi			
Incidence and risk of sorafenib-induced hypertension: a systematic review and meta-an	Journal of Clinical Hype	Exclude	Exclude	Exclude	Exclude	Exclu			
Increase of physical activity in essential hypertension - rapid report (Structured abstrac	Health Technology Ass	Include	Missing	Missing	Missing	Missi			
Increased risk of high-grade hypertension with bevacizumab in cancer patients: a meta	American Journal of Hy	Exclude	Exclude	Exclude	Exclude	Exclu			
Increasing physical activity for the treatment of hypertension: a systematic review and i	Sports Medicine	Include	Include	Missing	Include	Includ			

GreaseMonkey scripts used for screen scraping databases to acquire fuller citation details

GreaseMonkey - Epistemonikos download

```
// ==UserScript==
// @name Epistemonikos Download All
// @namespace http://crebp.net.au
// @version 1.0
// @description Script to automatically save all results in a Epistemonikos search to an RIS file.
// @include http://www.epistemonikos.org/en/search?*
// @include http://epistemonikos.org/en/search?*
// @grant none
// @require http://code.jquery.com/jquery-1.11.0.min.js
// @require http://raw.github.com/eligrey/FileSaver.js/master/FileSaver.js
// @require http://medialize.github.io/URI.js/src/URI.min.js
// @copyright 2014+, Matt Carter <m@ttcarter.com>
// @downloadURL https://raw2.github.com/CREBP/GreaseMonkey/master/Epistemonikos%20Download%20All.js
// @updateURL https://raw2.github.com/CREBP/GreaseMonkey/master/Epistemonikos%20Download%20All.js
// ==/UserScript==
$(function() {
console.log('SDL', $('#selected_documents_link'))
$('#selected_documents_link > p').css('margin-bottom', '10px');
$('<a title="Download all references" class="btn btn-primary btn-sm pull-right" href="#"><i class="glyphicon glyphicon-download-alt"></i> Download All</a>')
.appendTo($('#selected_documents_link > p'))
.after(' ')
.on('click', function() {
if ($.downloadAll && $.downloadAll.refs) { // Already done the work
$.downloadAll.generateOutput();
return;
}
$('#modal-da-progress').remove();
$('body').append('<div id="modal-da-progress" class="modal">' +
'<div class="modal-dialog"><div class="modal-content">' +
'<div class="modal-header">' +
'<button type="button" class="close" data-dismiss="modal" aria-hidden="true">&times;</button>' +
'<h3>Processing references...</h3>' +
'</div>' +
'<div class="modal-body">' +
```

```
'<div class="progress progress-striped active">' +
'<div id="modal-da-progress-bar" class="progress-bar" role="progressbar" aria-
valuenow="0" aria-valuemin="0" aria-valuemax="100" style="width: 0%"></div>' +
'</div>' +
'<p id="modal-da-progress-text" class="text-center">Preparing...</p>' +
'</div>' +
'<div class="modal-footer">' +
'<a href="#" class="btn btn-danger" data-dismiss="modal">Cancel</a>' +
'</div>' +
'</div></div>' +
'</div>');
$.downloadAll = {
pageLink: $('.pagination .next').attr('href').replace(/p=([0-9]+)/, 'p=0'),
pageCurrent: 0,
pageCount: $('.pagination > li > a').not('.next').last().text(),
refs: [],
pageDownload: function() {
$.ajax({
url: $.downloadAll.pageLink.replace(/p=([0-9]+)/, 'p=' +
$.downloadAll.pageCurrent),
dataType: 'html',
error: function(err,txt) {
alert('An error has occured: ' + txt);
$('#modal-da-progress').hide();
},
success: function(data) {
if (!$.downloadAll) // Cancelled?
return;
$('#modal-da-progress-bar').css('width', parseInt(($.downloadAll.pageCurrent /
$.downloadAll.pageCount) * 100) + '%');
$('#modal-da-progress-text').text('Processing page ' + $.downloadAll.pageCurrent
+ ' of ' + $.downloadAll.pageCount);
$(data).find('.result').each(function() {
var me = $(this);
$.downloadAll.refs.push({
url: 'http://www.epistemonikos.org/' + me.find('h3 > a').attr('href'),
title: me.find('h3 > a').text(),
authors: me.find('.result-metadata > .authors .author').map(function() { return
$(this).text() }),
journal: me.find('.result-metadata > #journal > span').last().text(),
version: me.find('.result-metadata > #year > span').last().text()
});
});
if (++$.downloadAll.pageCurrent > $.downloadAll.pageCount) {
$('#modal-da-progress-bar').css('width', '100%');
$('#modal-da-progress-text').text('Compiling results');
```

```
setTimeout(function() {
$('#modal-da-progress').hide();
}, 2000);
$.downloadAll.generateOutput();
} else {
$.downloadAll.pageDownload();
}
}
});
},
cancel: function() {
$.downloadAll = null;
$('#modal-da-progress').hide();
},
translateVersion: function(str) {
var months =
{'January':1,'February':2,'March':3,'April':4,'May':5,'June':6,'July':7,'August':8,'Sept
ember':9,'October':10,'November':11,'December':12};
var matches = /^([A-Z][a-z]+) ([0-9]+)$/.exec(str);
if (matches)
return matches[2] + '/' + months[matches[1]] + '//';
matches = /^([0-9]+)$/.exec(str);
if (matches)
return matches[1] + '///';
return '';
},
generateOutput: function() {
var out = [];
for (var x = 0; x < $.downloadAll.refs.length; x++) {
var info = "TY - ELEC\n";
for (var a = 0; a < $.downloadAll.refs[x].authors.length; a++)
info += "AU - " + $.downloadAll.refs[x].authors[a] + "\n";
info +=
"PY - " + $.downloadAll.translateVersion($.downloadAll.refs[x].version) + "\n" +
"TI - " + $.downloadAll.refs[x].title + "\n" +
"JO - " + $.downloadAll.refs[x].journal + "\n" +
"DO - " + $.downloadAll.refs[x].url + "\n" +
"ER - \n";
out.push(info);
}
var blob = new Blob(out, {type: "text/plain;charset=utf-8"});
saveAs(blob, "Epistemonikos.ris");
}
};
$('#modal-da-progress')
.on('hide.bs.modal', $.downloadAll.cancel)
```

```
.on('click', '[data-dismiss="modal"]', $.downloadAll.cancel)
.show();
$.downloadAll.pageDownload(); // Start everything
```

GreaseMonkey - PubMed Health download

```
// ==UserScript==
// @name PubMed Health Download All
// @namespace http://crebp.net.au
// @version 1.0
// @description Script to automatically save all results in a PubMed Health search
to an RIS file.
// @include http://www.ncbi.nlm.nih.gov/pubmedhealth/*
// @include https://www.ncbi.nlm.nih.gov/pubmedhealth/*
// @grant none
// @require http://code.jquery.com/jquery-1.11.0.min.js
// @require http://raw.github.com/eligrey/FileSaver.js/master/FileSaver.js
// @require http://medialize.github.io/URI.js/src/URI.min.js
// @require http://netdna.bootstrapcdn.com/bootstrap/3.1.1/js/bootstrap.min.js
// @copyright 2014+, Matt Carter <m@ttcarter.com>
// @downloadURL
https://raw2.github.com/CREBP/GreaseMonkey/master/PubMed%20Health%20
Download%20All.js
// @updateURL
https://raw2.github.com/CREBP/GreaseMonkey/master/PubMed%20Health%20
Download%20All.js
// ==/UserScript==
$(function() {
$('body')
.prepend('<link rel="stylesheet"
href="http://netdna.bootstrapcdn.com/bootstrap/3.1.1/css/bootstrap.min.css"
type="text/css"/>')
.css('font-size', '10px');
if (window.location.href.substr(0, 8) == 'https://') // Switch to http:// version
window.location.href = window.location.href.replace('https://', 'http://');
$('a[data-value_id]').on('click', function() { // Fix the stupid inline link filter thats
used on the site
var myURI = URI(window.location)
.setSearch('filters', $(this).data('value_id'));
window.location.replace(myURI.toString());
});
$('<a title="Download all references" class="active page_link" href="#">Download
All</a>')
.prependTo($('.pagination'))
.after(' ')
.on('click', function() {
if ($.downloadAll && $.downloadAll.refs) { // Already done the work
$.downloadAll.generateOutput();
return;
}
```

```
$('#modal-da-progress').remove();
$('body').append('<div id="modal-da-progress" class="modal fade">' +
'<div class="modal-dialog"><div class="modal-content">' +
'<div class="modal-header">' +
'<button type="button" class="close" data-dismiss="modal" aria-
hidden="true">&times;</button>' +
'<h3>Processing references...</h3>' +
'</div>' +
'<div class="modal-body">' +
'<div class="progress progress-striped active">' +
'<div id="modal-da-progress-bar" class="progress-bar" role="progressbar" aria-
valuenow="0" aria-valuemin="0" aria-valuemax="100" style="width: 0%"></div>' +
'</div>' +
'<p id="modal-da-progress-text" class="text-center">Preparing...</p>' +
'</div>' +
'<div class="modal-footer">' +
'<a href="#" class="btn btn-danger" data-dismiss="modal">Cancel</a>' +
'</div>' +
'</div></div>' +
'</div>');
$.downloadAll = {
pageLink: $('.page_link.next').attr('href').replace(/page=([0-9]+)/, 'page=1'),
pageCurrent: 1,
pageCount: $('.pagination .page_link:last').attr('page'),
refs: [],
pageDownload: function() {
$.ajax({
url: $.downloadAll.pageLink.replace(/page=([0-9]+)/, 'page=' +
$.downloadAll.pageCurrent),
dataType: 'html',
error: function(err,txt) {
alert('An error has occured: ' + txt);
$('#modal-da-progress').modal('hide');
},
success: function(data) {
if (!$.downloadAll) // Cancelled?
return;
$('#modal-da-progress-bar').css('width', parseInt(($.downloadAll.pageCurrent /
$.downloadAll.pageCount) * 100) + '%');
$('#modal-da-progress-text').text('Processing page ' + $.downloadAll.pageCurrent
+ ' of ' + $.downloadAll.pageCount);
$(data).find('.rprt > .rslt').each(function() {
var me = $(this);
$.downloadAll.refs.push({
url: 'http://www.ncbi.nlm.nih.gov/' + me.find('.title > a').attr('href'),
title: me.find('.title > a').text(),
```

```
author: me.find('.supp > .details').text(),
version: me.find('.rprtid').text()
});
});
if (++$.downloadAll.pageCurrent > $.downloadAll.pageCount) {
$('#modal-da-progress-bar').css('width', '100%');
$('#modal-da-progress-text').text('Compiling results');
setTimeout(function() {
$('#modal-da-progress').modal('hide');
}, 2000);
$.downloadAll.generateOutput();
} else {
$.downloadAll.pageDownload();
}
}
});
},
cancel: function() {
$.downloadAll = null;
},
translateVersion: function(str) {
var months =
{'January':1,'February':2,'March':3,'April':4,'May':5,'June':6,'July':7,'August':8,'Sept
ember':9,'October':10,'November':11,'December':12};
var matches = /^Version: ([A-Z][a-z]+) ([0-9]+)$/.exec(str);
if (matches)
return matches[2] + '/' + months[matches[1]] + '//';
matches = /^Version: ([0-9]+)$/.exec(str);
if (matches)
return matches[1] + '///';
return '';
},
generateOutput: function() {
var out = [];
for (var x = 0; x < $.downloadAll.refs.length; x++) {
out.push(
"TY - ELEC\n" +
"AU - " + $.downloadAll.refs[x].author + "\n" +
"PY - " + $.downloadAll.translateVersion($.downloadAll.refs[x].version) + "\n" +
"TI - " + $.downloadAll.refs[x].title + "\n" +
"DO - " + $.downloadAll.refs[x].url + "\n" +
"ER - \n"
);
}
var blob = new Blob(out, {type: "text/plain;charset=utf-8"});
saveAs(blob, "PubMed Health.ris");
```

```
}
};
$('#modal-da-progress')
.on('hide.bs.modal', $.downloadAll.cancel)
.modal('show');
$.downloadAll.pageDownload(); // Start everything
});
});
```

GreaseMonkey - TRIP download

```
// ==UserScript==
// @name TripDatabase.com Download All
// @namespace http://crebp.net.au
// @version 1.0
// @description Script to automatically save all results in a TripDatabase search to an RIS file.
// @include http://www.tripdatabase.com/search?*
// @include http://tripdatabase.com/search?*
// @grant none
// @require http://medialize.github.io/URI.js/src/URI.min.js
// @copyright 2014+, Matt Carter <m@ttcarter.com>
// @downloadURL https://raw2.github.com/CREBP/GreaseMonkey/master/TripDatabase.com%20Download%20All.js
// @updateURL https://raw2.github.com/CREBP/GreaseMonkey/master/TripDatabase.com%20Download%20All.js
// ==/UserScript==
$(function() {
$('<a class="btn"><i class="icon icon-download" style="font-size: 13px"></i> Download All</a>')
.appendTo($('#results .results-meta'))
.before(' ')
.on('click', function() {
var myURI = URI(window.location)
.addSearch('max', '999999')
.path('/search/ris');
window.location.replace(myURI.toString());
});
});
```

Supplementary appendix B Deduplication

Additional methods

Creation of an Endnote library

In Endnote (X6, Thomson Reuters, USA), a library was created by selecting columns to display relevant fields to aid with the identification of unique and duplicate records. The fields chosen were: Author, Year, Title, Journal, Volume, Page number, Caption (used for record ID), and Notes and Language fields. The Notes field was used to enter coding for the benchmark ('Gold standard') to compare the performance of Endnote and the SRA-DM algorithm according to whether the record was unique or a duplicate, and whether the assigned duplicate record was identified correctly i.e. a genuine duplicate or a wrongly identified duplicate. The Language field was used to populate the decisions made by the SRA-DM algorithm.

Definitions

A duplicate record was defined as being the same bibliographic record (irrespective of how the citation details were reported, e.g. variations in page numbers, author details, accents used, or abridged titles). Where data from a single study was reported in several publications, these were not classed as duplicates, as they are multiple reports which can appear across or within journals. Similarly, where different publications report the *exact* data, e.g. journal and conference proceedings, these were treated as separate bibliographic records.

Benchmark

A dataset of 1988 records derived from a search conducted on 29th July 2013 for surgical and non-surgical management for pleural empyema was used to evaluate Endnote and SRA-DM algorithm. Six databases were searched (Box 1).

Box 1. ***Search dates***

- **Medline (Ovid) searched from 1946 to July week 3 2013**

  EMBASE searched from 2010 to July 2013 (all EMBASE RCTs prior to 2010 are now in CENTRAL)

- **CENTRAL searched July 2013 Issue 7**

- **CINAHL searched from 1981 to July 2013**

- **LILACS searched from 1982 to July 2013**

- **PUBMED searched July 2013 (searched to find any citations not listed in MEDLINE)**

Coding citations: identified as duplicates by Endnote

The 1988 records were imported into Endnote library and duplicate records were sought using the automated '*Find Duplicates*' command in Endnote. The default Endnote setting ('author', 'year', 'title') was used to match the records. The records returned as duplicates were visually inspected to check the accuracy, i.e. that all contained at least one duplicate. These records were coded in the Notes field as either Endnote Main (EM) duplicate, or Endnote True (ET) duplicate, i.e. ET being a true duplicate of EM. The first duplicate record to be identified (alphabetically sorted by author) was designated EM and its associated duplicate record(s) were coded as ET. Where records were incorrectly identified as being a duplicate, these were

coded as (EF), i.e. Endnote False duplicate.

Coding citations: identified as NOT duplicates by Endnote

The remaining records (not identified by Endnote as duplicates) were alphabetically ordered according to first author's name and then visually inspected to identify additional duplicate records. These records were coded in the Notes field as follows: if a record was identified as a *further* duplicate of EM, it was coded (DEM) i.e. Duplicate of Endnote Main; unique records were coded as (U). Duplicates existing only in the remaining records and not related to those duplicates identified by Endnote were coded as Original Duplicate (OD) and DOD, i.e. a Duplicate of OD.

Duplicate records could potentially go undetected in scenarios where first author names are misspelt or reported in a different order. Therefore, to avoid missed duplicates, the database was also alphabetically re-ordered according to the study titles and reinspected for duplicate records, before finalising the classification of records. This thorough process of labelling and double-checking each record reduced the likelihood of missing duplicates whilst establishing a benchmark against which the results of Endnote and the SRA-DM algorithm could be compared, and any anomalies could be pinpointed down to an individual record level. Therefore, each record was classified into one of the seven possible categories (Table 1).

Table 1. Individual record coding and definitions

Coding	Classification
EM	Endnote Main duplicate (identified correctly as a duplicate by Endnote)
ET	Endnote True duplicate of EM (identified correctly as an associated duplicate by Endnote)
EF	Endnote False duplicate of EM (incorrectly identified as a duplicate by Endnote)
U	Unique reference
DEM	Duplicate of EM (duplicate missed by Endnote)
OD	Original Duplicate (duplicate missed by Endnote)
DOD	Duplicate of OD (duplicate missed by Endnote)

Smart group filtering

To assess the performance of Endnote deduplication, the smart groups filtering facility within Endnote was used to further classify the 7 types of records into 4 main groups (Table 2): True Positives (EM and ET), False Positives (EF), True Negatives (U), and False Negatives (OD, DOD and DEM). These values were used to generate the sensitivity and specificity.

Table 2. Smart Group coding and definition

Group	Group characteristic
True Positive TP	True duplicate or correctly identified duplicate (EM or ET)
False Positive FP	False duplicate or incorrectly identified duplicate (EF)
True Negative TN	Unique record, correctly identified as non-duplicate (U)
False Negative FN	True duplicate, incorrectly identified as non-duplicate (OD, DOD or DEM)

Coding citations identified as duplicates by SRA

The SRA-DM algorithm identified and coded duplicate records automatically using a binary code - either 'OK' for either a unique record, or a record identified as a first duplicate, or 'DUPE' where the second or multiple duplicates were identified. The term 'OK' indicated that the record may be relevant and should be retained, whilst 'DUPE' signified a record that would be discarded.

The first step in validating the system was to filter out all 'OK' records designated by SRA-DM algorithm that corresponded to the unique (U) benchmark records, using the Create Smart Group facility in Endnote, and classifying them as True Negative (TN). This constituted the majority of the records. The remaining records were then visually and manually cross-checked against the benchmark coding to determine the accuracy of the SRA-DM algorithm.

Each record coded correctly as ‘OK’ corresponds to either a unique record (U), the main duplicate reference (EM) or a record incorrectly marked as duplicate by Endnote (EF), and these combinations were labelled as TN (True Negative). A correctly coded ‘DUPE’ corresponds to either the associated original duplicate records (OD) or to a redundant duplicate (ET, DEM or DOD). These were labelled TP (True Positive). The remaining records were marked as either FN or FP, as appropriate. Hence, where SRA-DM algorithm correctly matched according to the benchmark coding these were coded as either TP (True Positive), i.e. correctly identified duplicates, FP (False Positive), i.e. incorrectly identified duplicates, FN (False Negative), i.e. incorrectly identified unique records or TN (True Negative) for correctly identified unique records.

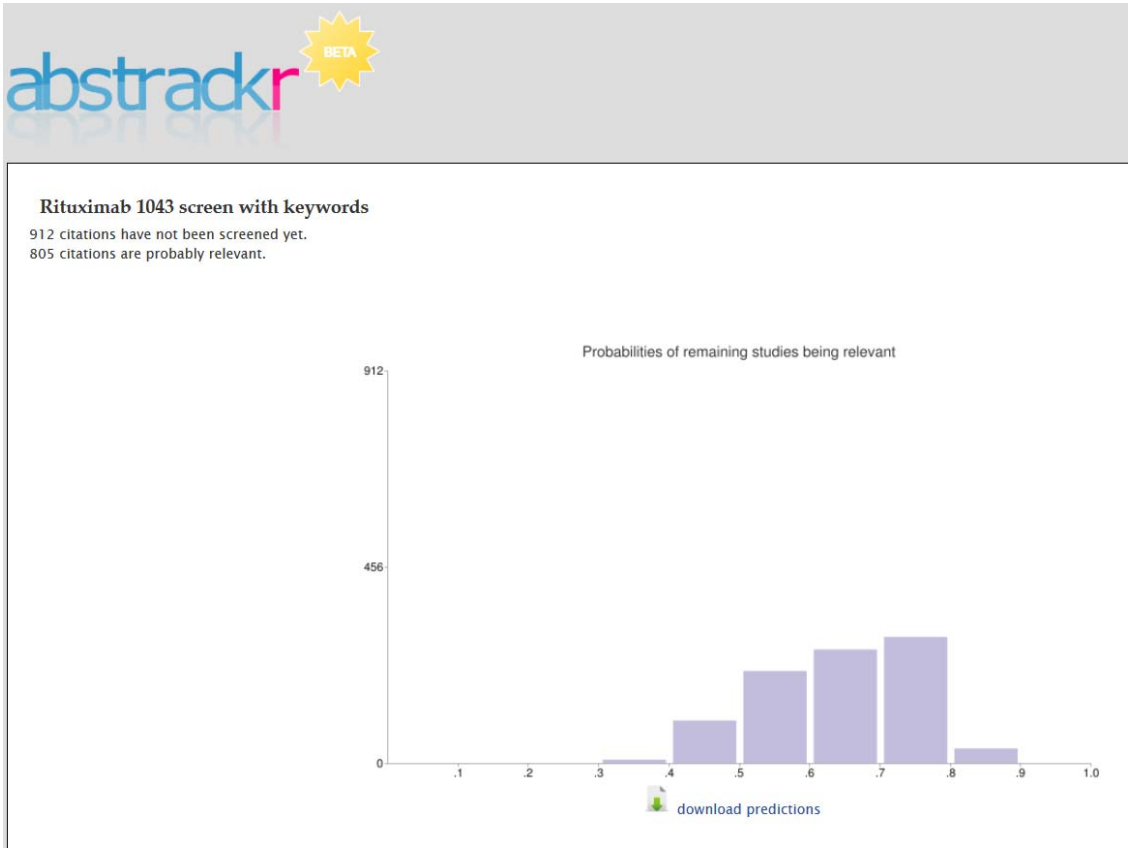
Supplementary appendix C Predictive screening (Abstrackr)

Example of training values (labels) applied to citations during the training phase in Abstrackr.

labels you've provided:

doc id	refman id	pubmed id	title	label
393298	291	0	Dexamethasone as an adjunctive treatment of bacterial meningitis	1
393102	539	0	Chemotherapy of tuberculosis meningitis with isoniazid plus rifampicin- interim findings in a trial in children [abstract]	-1
392733	139	0	[Immunoglobulins in the treatment of bacterial meningitis in childhood]	-1
393199	70	0	Trial of chloramphenicol for meningitis in northern savanna of Africa	1
392946	501	0	[Diffusion of amoxicillin and ampicillin intraenous administration in acute meningitis of children]	-1
393181	495	0	Gamma-Venin P in the treatment of bacterial meningitis in children	1
393315	5	0	Ampicillin compared with penicillin and chloramphenicol combined in the treatment of bacterial meningitis	1
392857	269	0	Dexamethasone and bacterial meningitis. A meta-analysis of randomized controlled trials	1
393266	497	0	Acute evoked potentials in patients with bacterial meningitis with and without Dexamethasonotherapy with regard to liquorelastase	1
393081	391	0	Randomised trial of Haemophilus influenzae type-b tetanus protein conjugate for prevention of pneumonia and meningitis in Gambian infants	-1

Example of Abstrackr prediction probabilities



Example of Abstrackr prediction probabilities and hard screening predictions

	A	B	C	D
1	citation_id	title	predicted p of being relevant	'hard' screening prediction*
2	430097	Mabthera (Rituximab) plus CVP chemotherapy for first-line treatment of stage III/IV follicular non-H	0.820092	TRUE
3	430086	Rituximab for treatment of intermediate and aggressive B-cell non-Hodgkin's lymphomas	0.81449	TRUE
4	429773	Effect of the addition of rituximab to front line therapy with cyclophosphamide, doxorubicin, vincri	0.81323	TRUE
5	429966	Treatment of relapsed B-cell non-Hodgkin's lymphoma with a combination of chimeric anti-CD20 m	0.81302	TRUE
6	429812	The effect of Rituximab on patients with follicular and mantle-cell lymphoma	0.812242	TRUE
7	430601	Front - Line Combined Immuno-Chemotherapy (R-CHOP) Significantly Improves the Time to Treatm	0.812236	TRUE
8	430651	Rituximab for treatment of intermediate and aggressive B-cell non-Hodgkin's lymphomas (Provisio	0.812172	TRUE
9	429753	The addition of Rituximab to a Fludarabine combination results in superior remission and survival r	0.812032	TRUE
10	429935	Phase II Randomized Study of Pegfilgrastim For Neutropenia After Cyclophosphamide, Doxorubicin	0.811347	TRUE
11	430111	Randomized comparison of MACOP-B with CHOP in patients with intermediate-grade non-Hodgkin	0.809966	TRUE
12	429868	Phase 2 Study of Bortezomib Weekly or Twice Weekly Plus Rituximab in Patients with Follicular (FL)	0.809126	TRUE
13	430005	Chimeric anti-CD20 monoclonal antibody (Rituximab; Mabthera) in remission induction and mainte	0.808501	TRUE
14	430612	Rituximab added to first-line mitoxantrone, chlorambucil, and prednisolone chemotherapy followe	0.808418	TRUE
15	429943	Primary diffuse large B-cell lymphoma of the testis (PTL): A prospective study of rituximab (R)-CHO	0.807952	TRUE
16	430018	CEOP-21 versus CEOP-14 chemotherapy with or without rituximab for the first-line treatment of pat	0.807727	TRUE
17	430277	Reduction of tumor burden and stabilization of disease by systemic therapy with anti-CD20 antibod	0.807393	TRUE
18	430163	Prolonged treatment with rituximab significantly improves event free survival and duration of resp	0.807378	TRUE
19	430483	Rituximab anti-CD20 monoclonal antibody therapy in non-Hodgkin's lymphoma: safety and efficacy	0.806188	TRUE
20	430655	Intralesional therapy with anti-CD20 monoclonal antibody rituximab in primary cutaneous B-cell lyr	0.805569	TRUE
21	430332	Optimizing the use of rituximab for treatment of B-cell non-Hodgkin's lymphoma: a benefit-risk up	0.804852	TRUE
22	429691	Chimeric anti-CD20 monoclonal antibody (rituximab; mabthera) in remission induction and mainter	0.804681	TRUE
23	429832	Treatment of relapsed B-cell non Hodgkin's lymphoma with a combination of chimeric anti-CD20 m	0.803775	TRUE
24	429792	Fludarabine plus mitoxantrone with and without rituximab versus CHOP with and without rituxima	0.803619	TRUE
25	429820	Rituximab combined to ACVBP (R-ACVBP) as a new inductive treatment followed by high-dose cons	0.803187	TRUE
26	430384	Anti-CD20 monoclonal antibody (Rituximab) in gastric extranodal marginal zone (MALT) non-Hodgk	0.802981	TRUE
27	430136	First-line and maintenance treatment with rituximab for patients with indolent non-Hodgkin's lymp	0.802262	TRUE
28	430434	Health Outcomes and Costs of Rituximab in Combination with Cyclophosphamide, Vincristine and P	0.802063	TRUE
29	430309	Health Outcomes and Costs of Rituximab in Combination with Cyclophosphamide, Vincristine and P	0.802063	TRUE
30	430101	The addition of Rituximab to combination chemotherapy with CHOP has a long lasting impact on su	0.801981	TRUE
31	430480	Phase III Randomized Study of Rituximab and Pixantrone (BBR 2778) Versus Rituximab Alone in Pati	0.801973	TRUE
32	429997	Randomized trial of r-metHu granulocyte colony-stimulating factor (G-CSF) as adjunct to CHOP or CH	0.801901	TRUE
33	430183	Prolonged Treatment with Rituximab Significantly Improves Event Free Survival and Duration of Re	0.801491	TRUE
34	429736	Rituximab in combination with chop improves survival in elderly patients with aggressive non-Hod	0.800359	TRUE
35	430040	THE ADDITION OF RITUXIMAB TO A FLUDARABINE COMBINATION (R-FCM) SIGNIFICANTLY IMPROVES	0.800046	TRUE
36	430173	Clinical study of Rituximab combined with CHOP in treatment of B cell non-Hodgkin's lymphoma	0.799855	TRUE
37	429823	A quality-adjusted survival analysis (Q-TWiST) of rituximab plus CVP vs CVP alone in first-line treatr	0.799833	TRUE